CONTRIBUTIONS TO THE THEORY OF ACCIDENT PRONENESS

I. AN OPTIMISTIC MODEL OF THE CORRELATION BETWEEN LIGHT AND SEVERE ACCIDENTS

BY

GRACE E. BATES and JERZY NEYMAN

UNIVERSITY OF CALIFORNIA PUBLICATIONS IN STATISTICS
Volume 1, No. 9, pp. 215–254, 7 figures in text

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1952

CONTRIBUTIONS TO THE THEORY OF ACCIDENT PRONENESS

I. AN OPTIMISTIC MODEL OF THE CORRELATION BETWEEN LIGHT AND SEVERE ACCIDENTS

BY

GRACE E. BATES AND JERZY NEYMAN

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1952

University of California Publications in Statistics
Editors: J. Neyman, M. Loève, O. Struve

Volume 1, No. 9, pp. 215–254, 7 figures in text

Submitted by editors November 26, 1951
Issued April 30, 1952
Price, 50 cents

University of California Press
Berkeley and Los Angeles
California

Cambridge University Press
London, England

PRINTED IN THE UNITED STATES OF AMERICA

DEDICATORY FOREWORD

THIS PAPER IS REVERENTLY DEDICATED TO THE MEMORY OF
GEORGE UDNY YULE, 1871–1951

When this paper was half written, the authors learned of the death of George Udny Yule. His death closed the early epoch of the development of the theory of statistics—an epoch marked by the names of F. Y. Edgeworth, W. S. Gosset ("Student"), Major Greenwood, Karl Pearson, W. F. Sheppard, and Yule, himself.

The contributions of Yule were numerous and were concerned with a number of aspects of statistical research, frequently in advance of his contemporaries. For many years, Yule was best known as the author of the book, *An Introduction to the Theory of Statistics*. First published in 1911, this book has had fourteen English editions (since 1937, revised editions have appeared under the joint authorship of G. U. Yule and M. G. Kendall) and for a long time was the only worthwhile book on statistics; several translations have also been published.

In more recent times, owing to the number of entirely new developments, the relative importance of the *Introduction* decreased and the name of George Udny Yule, as its author, began to slip into oblivion. At the same time, however, his name began to appear in the literature in various other connections—particularly in connection with what is now known as the theory of stochastic processes. Although by-passing the *Introduction*, modern statistical thought eventually caught up with a number of fruitful ideas published by Yule in the 1920's. At the time these ideas went hardly noticed but now proved *aere perennius*. Yule's own attitude towards mathematical statistics was distinctly nonmathematical, and it is, therefore, remarkable that his nonmathematical writings should now become a source of inspiration in the mathematical theory of stochastic processes. To us, this is the finest possible testimony to Yule's great scientific talent, and it is hoped that the frequent references to Yule by such authors as William Feller, Hermann Wold, and others may have cheered the aged scholar during the last years of his life.

In 1931 Yule felt that he was too old to hold the position of Reader at Cambridge University and retired. At the same time he felt young enough to learn to fly. Accordingly, he went through the intricacies of training, got a pilot's license, and bought a plane. Unfortunately a heart attack cut short both the flying and, to a considerable degree, his scholarly work.

Most of Yule's active life (roughly from 1897 to 1938) coincided with a tumultuous period in the development of mathematical statistics, when true scholarly achievements were accompanied by outbursts of personal animosities, noisy self-glorifications, and bitter disputes. Yule was an active scholar and it was natural for him to be under attack from time to time. However, to our knowledge, nothing Yule ever wrote conflicted with the dignity of the spirit of research, and his name enters history unmarred.

The range of Yule's scientific contributions was very broad. Among other things, he did pioneering work on accident proneness, in collaboration with Greenwood. In fact, the first line of the Introduction of the present paper contains a reference to their fundamental memoir. It is fitting that this paper be dedicated to the memory of George Udney Yule.

CONTRIBUTIONS TO THE THEORY OF ACCIDENT PRONENESS

I. AN OPTIMISTIC MODEL OF THE CORRELATION BETWEEN LIGHT AND SEVERE ACCIDENTS

BY
GRACE E. BATES AND JERZY NEYMAN

1. Introduction. Since the pioneer work of Greenwood and Yule [**1**][1] and of Miss Newbold [**2**], the following assumptions regarding accident proneness are customarily made:

a) To each individual exposed to a certain system of risks and to each kind of accident there corresponds a Poisson frequency function,

$$p_X(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

of the number X of accidents of this particular kind incurred by this individual per unit time.

b) The value of the parameter λ varies from one individual of the population to another and characterizes his specific accident proneness.

c) More specifically, it is frequently assumed that for an individual randomly selected from a given population exposed to a fixed system of risks, the parameter λ is a particular value of a random variable Λ with probability density function

$$p_\Lambda(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \tag{2}$$

where the constants $\alpha > 0$ and $\beta > 0$ depend on the population considered and on the kind of accidents.

d) It is customary to assume that, although with the passing of time the value of λ corresponding to a given individual may change, this change is slight only and an individual who is particularly prone to accidents in his youth remains a bad risk more or less indefinitely.

The evidence in favor of (*a*), (*b*) and (*d*) frequently appears quite convincing. Therefore, in selecting personnel for certain hazardous occupations, attempts are made (Farmer and Chambers [**3**]) to eliminate individuals who are particularly accident prone by employing only those who in the past had no accidents of the particular kind under consideration or only a few such accidents. Also (Ove Lundberg [**4**]) attempts are made to use records of accidents sustained and of cases of illness to adjust the premiums in accident and health insurance to actual risks attached to particular individuals. In each instance, attention is directed to accidents or cases of sickness occurring in two different periods of time (past and future

This work, begun under contract with the School of Aviation Medicine, U.S. Air Force, was completed with the partial support of the Office of Naval Research. Dr. Bates, a member of the faculty of Mount Holyoke College, worked at the University of California on this project.

[1] Numbers in brackets refer to references at the end of the paper.

experience) but belonging to the same category. The problem studied is essentially whether or not the number of accidents of a specified kind observed in the past has a predictive value for the number of accidents of the same kind to be observed in the future.

This question is very relevant in many cases. However, for certain purposes it is not completely relevant and must be modified. Such, for example, is the case when it is desired to select appropriate personnel for highly hazardous occupations (for example, airplane pilots) where the first accident observed is frequently also the last. For this very reason, in selecting the personnel it is impracticable to judge the individuals on their past experience with respect to the particular severe accidents and the most one can do is to see whether or not the frequency of mild accidents incurred in the past is relevant from the point of view of severe accidents to which the individual may be exposed in the future.

Pursuing this direction of thought we shall study not one but two (or more; further generalization is immediate) random variables, say X and Y, representing the numbers of accidents incurred by the same individual, either within the same period of time or in two different periods. The variable Y will mean the number of "predictor" accidents, which we may hope to be able to observe prior to the decision of whether or not the given individual is suitable for the particular employment. On the other hand, the random variable X will be interpreted as the number of severe accidents to be observed in the future.

As in the theory of Greenwood, Yule, and Newbold, we shall postulate that, for each individual, the variables X and Y are independent and follow two distinct Poisson distributions with parameters λ and μ which characterize the proneness of this individual to the two kinds of accidents. Furthermore, we shall postulate that the values of λ and μ vary from one individual to another.

In order that the value of Y can serve as a predictor regarding the value of λ it is necessary that λ and μ be correlated in the population considered and the closer the correlation, the greater the value of Y as a predictor. Whether or not the constants λ and μ, corresponding to two different kinds of accidents, are closely correlated is a question of fact and can be answered only by using appropriate empirical data.

The main purpose of the present paper is to study the distribution of X and Y on the following somewhat far-reaching hypothesis. This hypothesis will be frequently referred to in this paper so it will be conveniently labeled the *fundamental hypothesis*. It involves two assumptions:

i) the expectation μ of the number of predictor accidents is a fixed multiple of the expectation λ of the number of severe accidents, $\mu = a\lambda$, where a is a constant;

ii) in the population studied the distribution of Λ follows the Pearson type III law assumed by Greenwood, Yule and Newbold, as described in (*c*) above.

It will be seen that assumption (*i*) is very strong and, *a priori*, one is inclined to doubt whether it could ever be exactly satisfied. Surprisingly enough, the theoretical joint distribution deduced from the fundamental hypothesis was found to give a satisfactory fit to several empirical distributions. It follows, then, that the measures of success of the selection for small values of λ using Y as predictor, deduced in this paper, may not be far off in relation to real practical problems. Needless to say,

practical applications of these formulae must be preceded by an empirical test of the validity of the model studied with respect to the particular accidents which may come under consideration.

Assumption (*ii*) is also very strong. However, any other assumption specifying the distribution of Λ would be equally strong but, if one wants to obtain numerically a frequency function of X and Y, it is unavoidable to ascribe a definite form to the distribution of Λ. The adoption of the Pearson type III law is justified both by its flexibility as an interpolation formula and by the tradition established by Greenwood, Yule, and Newbold. However, in the course of the study it appeared that some properties of the multivariate distribution of the numbers of accidents satisfying assumption (*i*) are independent of the actual form of the distribution of Λ. Also, they have an immediate bearing on the problem of selection of personnel and for these two reasons are particularly interesting.

Part II of the paper deals with the possibility of a deeper insight into the nature of the mechanism behind the observed frequency distribution of the number of accidents from one individual to another.

The specific problem considered is that of the distinction between the Greenwood–Yule–Newbold model described here and the model of Pólya (slightly generalized), assuming that the probabilities of accidents in a specified time interval not only vary with the duration of this time interval, but depend upon the number of accidents previously sustained ("contagion") and on the length of exposure to accidents which is interpreted as a measure of the experience gained in the particular kind of work.

The details of the plan of Part I of the paper are as follows.

In section 2, the problem of the joint distribution of severe and light accidents is considered in a form which is a little more general than that envisaged above. Assuming the fundamental hypothesis, we consider not two different kinds of accidents but an arbitrary number $s \geqq 2$, of which the first is treated as "severe accidents" and the remaining $s - 1$ as different kinds of light predictor accidents.

Let $X_1, X_2, \cdots, X_s$ be the numbers of accidents of each kind. It is found that these random variables follow a joint distribution which the authors do not remember having seen before and which they propose to term the multivariate negative binomial distribution. This distribution possesses several remarkable properties, similar to those of the multivariate normal distribution. The more important of these properties refer to any group of $m < s$ variables out of the s considered.

i) Whatever the group of m variables, for example, $X_1, X_2, \cdots, X_m$, the marginal joint distribution of this group is an m-variate negative binomial.

ii) The joint distribution of $X_1, X_2, \cdots, X_m$ and of the sum, say $\chi = X_{m+1} + X_{m+2} + \cdots + X_s$ is an $(m + 1)$-variate negative binomial distribution.

iii) The conditional joint distribution of $X_1, X_2, \cdots, X_m$, given that the other $s - m$ variables have assumed specified values, is an m-variate negative binomial distribution and depends only on the value x of the sum χ.

iv) The regression of X_1 on $X_{m+1}, X_{m+2}, \cdots, X_s$ is linear, for $m = 1, 2, \cdots, s - 1$.

Because of property (*iii*), the general case of $s - 1 \geqq 1$ kinds of light accidents reduces to the simplest case involving only two categories of accidents, severe acci-

dents and light accidents, with the latter category embracing all the $s - 1$ different categories of light accidents originally considered.

Section 3 contains formulae leading to the estimates of the parameters in the bivariate negative binomial distribution.

Section 4 is given to an empirical test of the fundamental hypothesis. As mentioned above, the basic idea is that, for particular individuals in a population, the expected number of light accidents in an earlier period is a fixed multiple of the expected number of severe accidents in a subsequent period. Unfortunately, no empirical data were available with which the authors could test directly whether or not it is safe to assume this. The best that could be done was to study certain analagous situations for which the data could be obtained. On the whole, the results of this empirical study are promising.

The fundamental hypothesis is tested on two sets of data, one of which is new. Because of the scarcity of published empirical material of this particular kind, the new data are reproduced in this paper in several tables which may be useful for further work.

Section 5 is given to the following practical question: assuming the admittedly far-reaching fundamental hypothesis regarding the close connection between light and severe accidents, what are the prospects of success in the selection of personnel using the records of light accidents? It is shown that, in certain cases at least, the effect of selection must be substantial.

Section 6 outlines methods to be used if and when data on light and on severe accidents are available. The study of severe accidents differs from that of light accidents by the fact that severe accidents are frequently not survived by the victims. Consequently, even if the model treated in this paper is strictly applicable to light and severe accidents, because of the distortions caused by fatal accidents, the joint distribution of the numbers of light and severe accidents will not be the bivariate negative binomial. Therefore, any empirical study relating to light and severe accidents will require an appropriate distribution. Such a distribution, based on the assumption that the probability of surviving an accident is constant, is given in section 6.

Throughout the paper the notation adopted is that of J. Neyman's recent book [7].

2. Multivariate distribution of the numbers of accidents. The subject of this section is the joint distribution of an arbitrary number s of random variables X_1, $X_2, \cdots, X_s$, where X_i represents the number of accidents of the ith kind, incurred by an individual randomly drawn from a population.

The method used is that of probability generating functions, introduced by Laplace. A modern presentation of the theory is given by Feller [6]. The probability generating function is defined for sets of random variables all capable of assuming only nonnegative integer values. It will be denoted by G with subscripts indicating the random variables to which it refers. The arguments of G will always be assumed not to exceed unity in absolute value so as to insure the convergence of the series representing G. When dealing with conditional distributions, the hypotheses on which these distributions are based will be symbolized to the right of the vertical bar that follows the arguments of the probability generating function. Thus the

conditional probability generating function of the random variables $X_1, X_2, \cdots, X_s$, given a hypothesis H will be denoted and defined as

$$(3) \qquad G_{X_1,X_2,\ldots,X_s}(u_1, u_2, \cdots, u_s|H) = E\left[\prod_{i=1}^{s} u_i^{X_i} \mid H\right]$$

$$= \sum_{n_1,n_2,\ldots,n_s,} P\{(X_1 = n_1)\,(X_2 = n_2) \cdots (X_s = n_s)\} \prod_{i=1}^{s} u_i^{n_i}$$

where the summation extends over all nonnegative values of each $n_i = 0, 1, 2, \cdots$, for $i = 1, 2, \cdots, s$.

In the following, we shall use several properties of probability generating functions which are direct consequences of the above definition.

Generalizing the conditions of the problem studied by Greenwood, Yule, and Newbold, we assume that to the population studied and the s different kinds of accidents there correspond s positive numbers $a_1 = 1, a_2, a_3, \cdots, a_s$. Thus, these numbers are the same for each individual of the population. We assume further that to every individual of the population there corresponds a positive number λ, measuring his particular proneness to accidents. For an individual to be randomly drawn from the population, this number is interpreted as a particular value of a random variable Λ. The distribution of Λ will be denoted by $F(\lambda) = P\{\Lambda \leqq \lambda\}$. Some of the results obtained are independent of any assumption regarding $F(\lambda)$ except that $F(0) = 0$ so that Λ is necessarily a positive random variable. However, most of the results are based on the assumption that the distribution function of Λ has the particular form postulated by Greenwood, Yule, and Newbold, representing the integral of the probability density (2).

Given a particular individual of the population, that is, given a fixed value of λ, we shall assume that the numbers of accidents $X_1, X_2, \cdots, X_s$ are mutually independent and that each follows a Poisson law with the expectation of X_i equal to $a_i\lambda$, $i = 1, 2, \cdots, s$. It follows that, given λ, the conditional joint probability generating function of $X_1, X_2, \cdots, X_s$ is

$$(4) \qquad G_{X_1,X_2,\ldots,X_s}(u_1, u_2, \cdots, u_s \mid \lambda) = \exp\left\{\lambda \sum_{i=1}^{s} a_i(u_i - 1)\right\}.$$

Replacing in (4) λ by the random variable Λ and taking the expectation with respect to the distribution of this variable, we obtain the absolute probability generating function,

$$(5) \qquad G_{X_1,X_2,\ldots,X_s}(u_1, u_2, \cdots, u_s) = E\,[G_{X_1,X_2,\ldots,X_s}(u_1, u_2, \cdots, u_s \mid \Lambda)]$$

$$= \int_0^\infty \exp\left\{\lambda \sum_{i=1}^{s} a_i(u_i - 1)\right\} dF\,(\lambda)$$

$$= \phi\left[\sum_{i=1}^{s} a_i(u_i - 1)\right],$$

where $\phi(t)$ is the Laplace transform of the distribution $F(\lambda)$,

$$\phi(t) = \int_0^\infty e^{t\lambda}\, dF(\lambda)\,. \tag{6}$$

It will be seen that for $t < 0$ the function $\phi(t)$ is indefinitely differentiable.

The Laplace transform of the distribution defined by (2) is, say

$$\phi^*(t) = \int_0^\infty e^{t\lambda}\, p_\Lambda(\lambda)\, d\lambda = \left[1 - \frac{t}{\beta}\right]^{-\alpha}. \tag{7}$$

Thus, on the assumption that (2) represents the probability density of Λ, the joint probability generating function of $X_1, X_2, \cdots, X_s$ is, say

$$G^*_{X_1, X_2, \ldots, X_s}(u_1, u_2, \cdots, u_s) = \left[1 + \sum_{i=1}^{s} b_i(1 - u_i)\right]^{-\alpha}, \tag{8}$$

where, for the sake of simplicity in formulae, $b_i = a_i/\beta$, $i = 1, 2, \cdots, s$.

Owing to the particular form of the probability generating function (8), the corresponding distribution of $X_1, X_2, \cdots, X_s$ will be called the s-variate negative binomial distribution. Easy expansion of (8) in powers of $u_1, u_2, \cdots, u_s$ gives

$$P\{(X_1 = n_1)\,(X_2 = n_2) \cdots (X_s = n_s)\} = \left[1 + \sum_{i=1}^{s} b_i\right]^{-\alpha} \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)} \prod_{i=1}^{s} \frac{c_i^{n_i}}{n_i!} \tag{9}$$

where $n = n_1 + n_2 + \cdots + n_s$ and

$$c_i = \frac{b_i}{1 + \sum_{j=1}^{s} b_j} = \frac{a_i}{\beta + \sum_{j=1}^{s} a_j}\,. \tag{10}$$

The distributions defined by (5) and (8) possess the following remarkable properties.

Let $r_1, r_2, \cdots, r_s$ be any permutation of numbers $1, 2, \cdots, s$ and let m be any positive integer less than s.

THEOREM 1. *If the random variables $X_1, X_2, \cdots, X_s$ follow the multivariate negative binomial distribution* (8) *then the joint distribution of $X_{r_1}, X_{r_2}, \cdots, X_{r_m}$ is also negative binomial.*

The probability generating function of the marginal distribution of $X_{r_1}, X_{r_2}, \cdots, X_{r_m}$ is obtained from (8) by substituting $u_{r_{m+1}} = u_{r_{m+2}} = \cdots = u_{r_s} = 1$. It is easily seen that the result of this substitution is a function of the same type with the sum of m terms

$$\sum_{i=1}^{m} b_{r_i}(1 - u_{r_i}) \tag{11}$$

replacing in the square brackets the sum of s similar terms and the theorem is proved.

THEOREM 2. *Whatever the distribution $F(\lambda)$ of Λ, given that the sum, say*

$$\chi = \sum_{i=1}^{s} X_i , \tag{12}$$

has assumed a value n, the conditional joint distribution of $X_1, X_2, \cdots X_{s-1}$ is the multinomial distribution with the probability generating function

$$G_{X_1,X_2,\ldots,X_{s-1}}(u_1, u_2, \cdots, u_{s-1} \mid \chi = n) = \left[\sum_{i=1}^{s-1} d_i u_i + d_s\right]^n , \tag{13}$$

with

$$d_i = \frac{a_i}{\sum_{j=1}^{s} a_j} , \qquad i = 1, 2, \cdots, s . \tag{14}$$

Starting with the definition, the generating function

$$\begin{aligned} G_{X_1,X_2,\ldots,X_{s-1},\chi}(u_1, u_2, \cdots, u_{s-1}, v) &= E\left[v^{\chi} \prod_{i=1}^{s-1} u_i^{X_i}\right] \\ &= E\left[v^{X_s} \prod_{i=1}^{s-1} (u_i v)^{X_i}\right] \\ &= G_{X_1,X_2,\ldots,X_{s-1},X_s}(u_1 v, u_2 v, \cdots, u_{s-1} v, v), \end{aligned} \tag{15}$$

and, therefore, because of (5)

$$G_{X_1,X_2,\ldots,X_{s-1},\chi}(u_1, u_2, \cdots, u_{s-1}, v) = \phi\left[v \sum_{i=1}^{s} a_i u_i - \sum_{i=1}^{s} a_i\right]\Bigg|_{u_s=1} . \tag{16}$$

In order to obtain the probability generating function (13) it is sufficient to expand (16) in powers of v, to select the coefficient of v^n and to divide this coefficient by its value corresponding to $u_1 = u_2 = \cdots = u_{s-1} = 1$. It is easy to see that the result of this operation coincides with (13).

THEOREM 3. *Whatever be the distribution function $F(\lambda)$ of Λ, given that $X_{r_i} = n_{r_i}$, $i = m + 1, m + 2, \cdots, s$, the conditional distribution of X_{r_j}, for $j = 1, 2, \cdots, m$, depends only on the sum*

$$n = \sum_{i=m+1}^{s} n_{r_i} \tag{17}$$

but not on the numbers n_{r_i} taken separately.

This theorem describes a very important property of the joint distribution of accidents satisfying assumption (*i*). Owing to this property, the problem of predicting the number of severe accidents, say, the number X_1 of accidents of the first kind, using the numbers, e.g., $X_{m+1}, X_{m+2}, \cdots, X_s$, of accidents of some $s - m$ other kinds reduces to that of predicting X_1 using the value of the sum, say

$$Y = \sum_{i=m+1}^{s} X_i \,. \tag{18}$$

Thus, whatever be the relative frequency of the predictor accidents as measured by the constants $a_{m+1}, a_{m+2}, \cdots, a_s$, in order to predict the value of X_1 no weighing of the numbers of these accidents is necessary, and this irrespective of the actual form of the distribution of Λ.

Obviously, it will be sufficient to prove theorem 3 for $r_i = i$, $i = 1, 2, \cdots, s$. By examining the definition of the probability generating function it is easy to see that the conditional probability generating function of $X_1, X_2, \cdots, X_m$ given that the other variables $X_{m+1}, X_{m+2}, \cdots, X_s$ have assumed some specified values $n_{m+1}, n_{m+2}, \cdots, n_s$, respectively, is obtained from (5) as a result of the two following operations.

a) Expand (5) in powers of $u_{m+1}, u_{m+2}, \cdots, u_s$ and obtain the coefficient C of the product

$$\prod_{i=m+1}^{s} u_i^{n_i} \,. \tag{19}$$

Obviously, C is a function of $u_1, u_2, \cdots, u_m$.

b) Divide C by the value of this coefficient corresponding to $u_1 = u_2 = \cdots = u_m = 1$.

Performing these operations on (5), we obtain

$$C = \phi^{(n)}(t) \prod_{i=m+1}^{s} \frac{a_i^{n_i}}{n_i!} \tag{20}$$

where $\phi^{(n)}(t)$ denotes the nth derivative of ϕ with respect to t and where

$$t = \sum_{j=1}^{m} a_i(u_i - 1) - \sum_{j=m+1}^{s} a_j = \sum_{i=1}^{m} a_i(u_i - 1) + \tau \,, \quad \text{say} \,. \tag{21}$$

It follows that

$$G_{X_1, X_2, \ldots, X_m}[u_1, u_2, \cdots, u_m \mid (X_{m+1} = n_{m+1}) \cdots (X_s = n_s)] = \frac{\phi^{(n)}(t)}{\phi^{(n)}(\tau)} \,. \tag{22}$$

It is seen that the right-hand side depends on the sum n of the values assumed by the variables $X_{m+1}, X_{m+2}, \cdots, X_s$ but not on these values taken separately, which proves theorem 3.

THEOREM 4. *If the variables $X_1, X_2, \cdots, X_s$ follow the multivariate negative binomial distribution* (8) *then, given $X_{r_i} = n_{r_i}$ for $i = m + 1, m + 2, \cdots, s$, the conditional distribution of $X_{r_1}, X_{r_2}, \cdots, X_{r_m}$ is also a negative binomial distribution depending on the sum n defined by* (17).

It will be sufficient to prove theorem 4 assuming $r_i = i$ for $i = 1, 2, \cdots, s$. The proof may either be direct, starting from (8) or take into account (22) and evaluate the nth derivative of (7). We have

$$\frac{d^n\phi^*}{dt^n} = \frac{1}{\beta^n}\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\left[1 - \frac{t}{\beta}\right]^{-(\alpha+n)} \tag{23}$$

and it follows that

$$G_{X_1,X_2,\ldots,X_m}[u_1, u_2, \cdots, u_m \mid (X_{m+1} = n_{m+1}) \cdots (X_s = n_s)] \tag{24}$$

$$= \left[\frac{\beta - t}{\beta - \tau}\right]^{-(\alpha+n)}$$

$$= \left[1 + \sum_{i=1}^{m} e_i(1 - u_i)\right]^{-(\alpha+n)}$$

with

$$e_i = \frac{a_i}{\beta + \sum_{j=m+1}^{s} a_j} \tag{25}$$

which proves the theorem.

As a result of theorem 3, the conditional distribution of $X_1, X_2, \cdots, X_m$, given $X_{m+1} = n_{m+1}, \cdots, X_s = n_s$, will be identified with the conditional distribution of the same variables, given that the sum Y defined in (18) has assumed the value n of (17). In particular, the multiple correlation coefficient of X_1 and $X_{m+1}, X_{m+2}, \cdots, X_s$, say ρ, coincides with the ordinary correlation coefficient of X_1 and Y as defined in (18). In order to study the regression of X_1 on $X_{m+1}, X_{m+2}, \cdots, X_s$ or the multiple correlation ρ, it will be sufficient to consider the probability generating function of X_1 and Y obtainable either from (5) or from (8) by substituting $u_1 = u$, $u_2 = u_3 = \cdots = u_m = 1$ and $u_{m+1} = u_{m+2} = \cdots = u_s = v$. Thus formula (5) gives

$$G_{X_1,Y}(u, v) = \phi[a_1(u - 1) + A(v - 1)] \tag{26}$$

where, for short,

$$A = \sum_{i=m+1}^{s} a_i \,. \tag{27}$$

THEOREM 5. *Whatever the distribution function $F(\lambda)$ of Λ, provided it possesses two first moments, the square of the correlation coefficient ρ^2 between X_1 and Y is given by*

$$\rho^2 = \left[1 - \frac{E(X_1)}{\sigma^2_{X_1}}\right]\left[1 - \frac{E(Y)}{\sigma^2_Y}\right] = \left[1 + \frac{\mu_1}{a_1\sigma^2_\Lambda}\right]^{-1}\left[1 + \frac{\mu_1}{A\sigma^2_\Lambda}\right]^{-1} \tag{28}$$

where μ_1 is the expectation of Λ and σ^2_Λ its variance.

In order to deduce formula (28) we use the familiar relations between the moments

of the random variables and the derivatives of their probability generating function evaluated at the values of arguments equal to unity. In particular

$$E(X_1) = \left.\frac{\partial G_{X_1,Y}}{\partial u}\right|_{u=v=1} = a_1\phi'(0) = a_1\mu_1 , \tag{29}$$

$$E(X_1^2) - E(X_1) = \left.\frac{\partial^2 G_{X_1,Y}}{\partial u^2}\right|_{u=v=1} = a_1^2\phi''(0) = a_1^2 E(\Lambda^2) \tag{30}$$

and it follows

$$\sigma_{X_1}^2 = a_1^2\sigma_\Lambda^2 + a_1\mu_1 . \tag{31}$$

Also,

$$E(Y) = A\mu_1 , \qquad \sigma_Y^2 = A^2\sigma_\Lambda^2 + A\mu_1 \tag{32}$$

and

$$E(X_1, Y) = \left.\frac{\partial^2 G_{X_1,Y}}{\partial u\,\partial v}\right|_{u=v=1} = a_1 A\, E(\Lambda^2) . \tag{33}$$

Finally, we get

$$\rho^2 = \frac{[E(X_1,\ Y) - E(X_1)\,E(Y)]^2}{\sigma_{X_1}^2\,\sigma_Y^2} = \frac{a_1 A\sigma_\Lambda^4}{(a_1\sigma_\Lambda^2 + \mu_1)\,(A\sigma_\Lambda^2 + \mu_1)} \tag{34}$$

which coincides with the second part of (28). In order to obtain the first part of this formula, we notice that

$$1 - \frac{E(X_1)}{\sigma_{X_1}^2} = \frac{a_1\sigma_\Lambda^2}{a_1\sigma_\Lambda^2 + \mu_1} \tag{35}$$

and a similar relation for Y.

Theorem 4 implies important conclusions regarding the possibility of predicting the value X_1 by using the values assumed by $X_2, X_3, \cdots, X_s$.

COROLLARY 1. *If ρ is taken as a conventional measure of precision in predicting the value of X_1 from the observed values of $X_{m+1}, X_{m+2}, \cdots, X_s$, then, whatever be $F(\lambda)$, it is advantageous to use as many predictors as possible, that is, it is advantageous to take $m = 1$.*

This conclusion is the immediate result of the fact that ρ is an increasing function of A as defined in (27).

COROLLARY 2. *Whatever the distribution function $F(\lambda)$, and whatever the number of predictor variables $X_{m+1}, X_{m+2}, \cdots, X_s$, the correlation ρ must be smaller than the upper bound*

$$\rho < \left[1 - \frac{E(X_1)}{\sigma_{X_1}^2}\right]^{\frac{1}{2}} \tag{36}$$

depending only on the expectation and on the variance of the predicted variable X_1.

Formula (36) is an immediate consequence of the first part of (28). The practical conclusion is that, before attempting to use the numbers of any accidents in order to predict the value of X_1, one should estimate the expectation of X_1 and its variance. If the right-hand side of (36) is close to zero then the prospects of attaining a good prediction, at least by means of a linear regression equation, are slim.

THEOREM 6. *If the random variables $X_1, X_2, \cdots, X_s$ jointly follow a multivariate negative binomial distribution, then the regression of X_1 on the sum Y as defined by* (18), *is linear, and namely*

$$E(X_1 \mid Y = n) = \frac{a_1(\alpha + n)}{\beta + \sum_{j=m+1}^{s} a_j} . \tag{37}$$

Under the hypotheses of the theorem, the conditional probability generating function of X_1, given $Y = n$, is obtained from (24) by substituting $u_2 = u_3 = \cdots = u_m = 1$ and we have

$$G_{X_1}(u_1 \mid Y = n) = [1 + e_1(1 - u_1)]^{-(\alpha+n)} . \tag{38}$$

The regression function of X_1 on Y is obtained by differentiating (38) and by setting $u_1 = 1$. The result is (37).

Theorems 1, 4, and 6 describe interesting properties of the multivariate negative binomial distribution whereby it is somewhat similar to the multivariate normal. Naturally, however, the analogy is far from complete. Thus, for example, the conditional variance of X_1 given Y, or given any single variable X_i, $i \neq 1$, is not constant but increases linearly with the value of the fixed variable. Furthermore, the sum of two independent negative binomial variables may but need not be a negative binomial, etc.

3. Estimation of parameters in the bivariate negative binomial distribution. In section 2 it was shown that, when the model considered applies, the s-dimensional problem may be reduced to a two-dimensional problem. In particular, if formula (2) adequately represents the probability density function of Λ, then, in order to treat the problem of predicting the number, say $X = x$, of severe accidents using any number of categories of light accidents, it is sufficient to study a bivariate negative binomial distribution of X and Y, where Y stands for the total number of light accidents embracing all the $s - 1$ different categories originally considered. Remembering the convention $a_1 = 1$, the joint probability generating function of X and Y may be written as

$$G_{X,Y}(u, v) = \frac{\beta^\alpha}{[\beta + (1 - u) + A(1 - v)]^\alpha} \tag{39}$$

with $A = a_2 + a_3 + \cdots + a_s$. By expanding (39) in powers of u and v, we obtain as the coefficient of $u^k v^m$

$$p_{X,Y}(k, m) = \frac{\Gamma(\alpha + m + k)}{k!\, m!\, \Gamma(\alpha)} \beta^\alpha A^m (\beta + A + 1)^{-(\alpha+m+k)} . \tag{40}$$

We shall suppose that n independent observations on the pair (X, Y) will be made. The letter $n_{k,m}$ will then denote the random variable representing the number of pairs $[(X = k), (Y = m)]$. The joint frequency function of all the $n_{k,m}$ is represented by the product, say

$$J = C \prod_{k=0}^{\infty} \prod_{m=0}^{\infty} p_{X,Y}^{n_{k,m}}(k, m) , \tag{41}$$

where C stands for a factor depending on the $n_{k,m}$ but not on the parameters α, β, and A, and where

$$\sum_{k=0}^{\infty} \sum_{m=0}^{\infty} n_{k,m} = n . \tag{42}$$

Our problem is to deduce formulae for the maximum likelihood estimates, say $\hat{\alpha}_n$, $\hat{\beta}_n$ and $\hat{A}_n$ of these three parameters. Recent results [**8**] imply that these estimates possess the following properties: (*i*) the estimates are functions of the relative frequencies $n_{k,m}/n$ but do not depend otherwise on n; (*ii*) the estimates possess continuous partial derivatives with respect to each relative frequency; (*iii*) as $n \to \infty$, the estimates are consistent and asymptotically normal about the true values of the particular parameters; (*iv*) the asymptotic variances of the estimates $\hat{\alpha}_n$, $\hat{\beta}_n$, and $\hat{A}_n$ decrease as n^{-1} and do not exceed the asymptotic variances of any other estimates possessing the properties (*i*), (*ii*) and (*iii*).[2]

Substituting (40) into (41), taking logarithms and dividing by n, we obtain

$$\frac{1}{n} \log J = C_1 + \alpha \log \beta - (\alpha + \bar{X} + \bar{Y}) \log (\beta + A + 1) + \bar{Y} \log A + \sum_{t=0}^{\infty} \left(1 - \sum_{r=0}^{t} q_r\right) \log (\alpha + t) \tag{43}$$

where C_1 represents a term independent of the parameters and where

$$\bar{X} = \frac{1}{n} \sum_{k=0}^{\infty} k \sum_{m=0}^{\infty} n_{k,m} ,$$

$$\bar{Y} = \frac{1}{n} \sum_{m=0}^{\infty} m \sum_{k=0}^{\infty} n_{k,m} , \tag{44}$$

$$q_r = \frac{1}{n} \sum_{k=0}^{r} n_{k,(r-k)} .$$

[2] Until recently it was believed that the asymptotic variances of the maximum likelihood estimates cannot exceed those of any other consistent and asymptotically normal estimates. A conjecture to this effect is usually ascribed to R. A. Fisher, who, since 1921 [**10**], has repeatedly proclaimed the above statement as a property of maximum likelihood estimates. In this connection, see also F. Y. Edgeworth who enunciated in his paper [**9**] of 1908 essentially the same conjecture (with a vague restriction on the nature of the estimate). Although the proofs of both Edgeworth and Fisher obviously lack precision, this conjecture was generally taken for granted and quoted in many articles and books. Recently J. L. Hodges, Jr. [**11**] has produced examples of consistent and asymptotically normal estimates, not having the properties (*i*) and (*ii*), whose asymptotic variances never exceed those of the maximum likelihood estimates and, for some values of the parameter, are actually smaller.

Obviously, q_r represents the relative frequency of pairs (X, Y) which have their sum $X + Y = r$. The maximum likelihood equations are obtained by differentiating (43) with respect to α, β and A and by equating the derivatives to zero. We have:

$$\log \frac{\hat{\beta}}{\hat{\beta} + \hat{A} + 1} + \sum_{t=0}^{\infty} \frac{1 - \sum_{r=0}^{t} q_r}{\hat{\alpha} + t} = 0 , \tag{45}$$

$$\frac{\hat{\alpha}}{\hat{\beta}} - \frac{\hat{\alpha} + \bar{X} + \bar{Y}}{\hat{\beta} + \hat{A} + 1} = 0 , \tag{46}$$

$$\frac{\bar{Y}}{\hat{A}} - \frac{\hat{\alpha} + \bar{X} + \bar{Y}}{\hat{\beta} + \hat{A} + 1} = 0 . \tag{47}$$

Equations (46) and (47) imply

$$\hat{\alpha} = \bar{X}\hat{\beta} , \tag{48}$$

$$\bar{Y} = \bar{X}\hat{A} , \tag{49}$$

and then equation (45) gives

$$\log \left(1 + \frac{\bar{X} + \bar{Y}}{\hat{\alpha}}\right) = \sum_{t=0}^{\infty} \frac{1 - \sum_{r=0}^{t} q_r}{\hat{\alpha} + t} . \tag{50}$$

The problem of computing the maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{A}$ is thus reduced to the following operations. First we calculate the means $\bar{X}$ and $\bar{Y}$ of the observed values of X and Y, respectively, and the relative frequencies q_r as indicated in formulae (44). Upon substituting them into (50) the trial and error method gives the value of $\hat{\alpha}$. Next

$$\hat{\beta} = \frac{\hat{\alpha}}{\bar{X}} , \qquad \hat{A} = \frac{\bar{Y}}{\bar{X}} . \tag{51}$$

In trying to obtain $\hat{\alpha}$ it is well to notice that the two sides of equation (50) tend to the same limit zero as α is indefinitely increased. The first trial value, say α_0, may be conveniently obtained as follows. We notice that the result of substituting $\hat{\alpha}$, $\hat{\beta}$ and $\hat{A}$ in

$$p_{X,Y}(0, 0) = \left(\frac{\beta}{\beta + A + 1}\right)^{\alpha} \tag{52}$$

should give a result comparable to q_0. Using equations (48) and (49) we have

$$\frac{\hat{\beta}}{\hat{\beta} + \hat{A} + 1} = \frac{\hat{\alpha}}{\hat{\alpha} + \bar{X} + \bar{Y}} . \tag{53}$$

Thus, the first trial value of $\hat{\alpha}$ can be taken to satisfy the equation

$$\left(\frac{\alpha_0}{\alpha_0 + \bar{X} + \bar{Y}}\right)^{\alpha_0} = q_0 \tag{54}$$

which is equivalent to

$$\log (1 + z) = -\frac{\log q_0}{\bar{X} + \bar{Y}} z \tag{55}$$

with $z = (\bar{X} + \bar{Y})/\alpha_0$. In order to obtain α_0 we make a graph of the logarithmic function

$$y = \log (1 + z) . \tag{56}$$

Next we plot the straight line

$$y = -\frac{\log q_0}{\bar{X} + \bar{Y}} z . \tag{57}$$

The two lines have two points in common, one at $z_1 = 0$ and the other at $z_2 = (\bar{X} + \bar{Y})/\alpha_0$, which is obtained graphically. When z_2 is obtained, we get $\alpha_0 = (\bar{X} + \bar{Y})/z_2$.

4. Empirical test of the fundamental hypothesis. As mentioned before, the validity of the fundamental hypothesis considered in this paper and, in particular, of the joint bivariate negative binomial distribution (40), should be tested with respect to the particular types of accidents that may come under study. Thus, for example, if it is attempted to apply the conclusions of this paper to the selection of airplane pilots through the use of an individual's record of minor accidents during the years before the training in order to obtain individuals with low proneness for aviation accidents, then the validity of the fundamental hypothesis should be tested on observations regarding the numbers X and Y of each kind of accident actually suffered by a number of individuals. Owing to the lack of data, no such test is possible at present. However, because of the far-reaching character of the fundamental hypothesis, it is of interest to inquire whether or not there are any accidents at all with respect to which this hypothesis is at least approximately true.

To investigate this point, formula (40) was tried in connection with the following two sets of data. The first set was obtained through the courtesy of Dr. Rosedith Sitgreaves and Dr. W. M. Gafafer, to whom the authors are deeply indebted. Special thanks are due to Dr. J. G. Townsend, Chief, Division of Industrial Hygiene, Public Health Service, Federal Security Agency, who released the data collected by the Division of Industrial Hygiene.

The data are concerned with two different categories of employees of an industrial establishment: Group 1 = office workers, and Group 2 = industrial workers. For each of these two groups the data list the numbers of cases of incapacity suffered during a period of time due to the following causes:

Cause 1 = Respiratory disease
Cause 2 = Digestive disease
Cause 3 = Nonindustrial injury
Cause 4 = Industrial injury

Each case of incapacity from any of the four causes was treated as an accident of a special category.

The other set of data on which the test of the fundamental hypothesis was made is taken from the publication of Farmer and Chambers [3]. This is concerned with accidents incurred by 166 London bus drivers during five successive years of service. On these data, two tests of the model were made, once taking the experience of the first four years of service of each man as one variable and the experience of the fifth as the other and then treating the number of accidents in the first year of service as one of the two variables and the number of accidents in the subsequent four years as the other.

TABLE 1

TEST OF THE VALIDITY OF THE FUNDAMENTAL HYPOTHESIS ON TWO SETS OF DATA

Data on:	Estimated parameters			No. of individuals	Degrees of freedom	$P(\chi^2)$
	α	β	A			
Employees of an industrial concern						
Cause 1 vs 2, Gr. 1.....	1.452	1.407	4.729	407	37	.10
Cause 1 vs 2, Gr. 2.....	1.471	1.050	3.798	1272	95	Practically zero
Cause 1 vs 3, Gr. 1.....	1.657	4.750	13.986	407	39	.090
Cause 1 vs 3, Gr. 2.....	1.686	4.734	15.075	1272	58	.00053
Cause 2 vs 3, Gr. 1.....	0.922	2.662	2.979	407	16	.0017
Cause 2 vs 3, Gr. 2.....	0.846	2.377	3.978	1272	30	Practically zero
Cause 3 vs 4, Gr. 1.....	1.309	28.046	8.421	407	3	.59
Cause 3 vs 4, Gr. 2.....	1.385	3.888	0.740	1272	11	Practically zero
London bus drivers						
Fifth year vs four first years.....	3.490	2.021	4.125	166	38	.35
First year vs last four years.....	5.596	3.086	3.419	166	32	.21

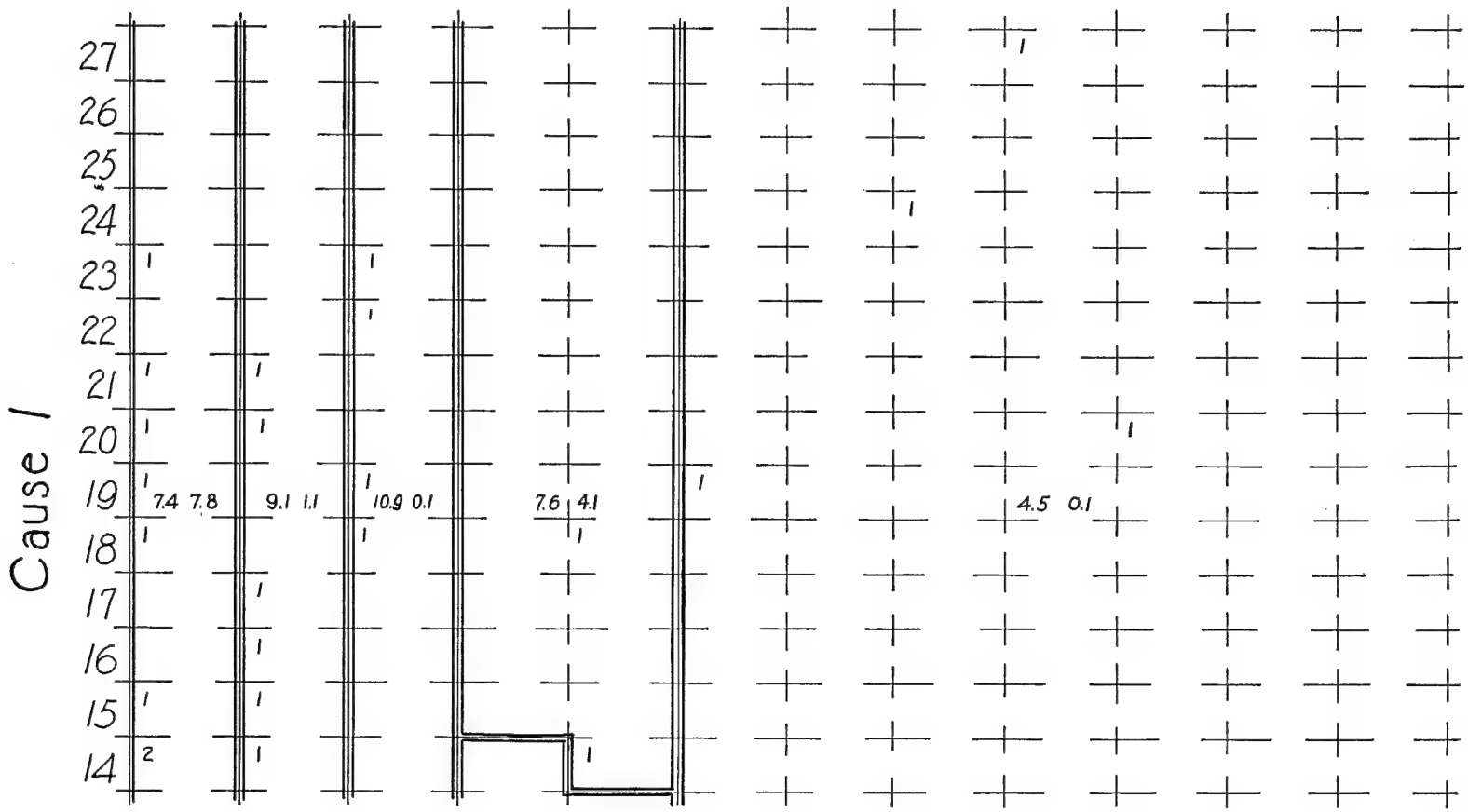
Table 1 gives the results of all these tests. The first three columns give the values of the estimated parameters of the distribution (40), the fourth column gives the number of individuals to whom the particular observations refer, the fifth the number of degrees of freedom in applying the χ^2 test and the sixth the value of the probability $P(\chi^2)$ of obtaining a value of χ^2 exceeding that observed.

Tables 2 to 11 give the bivariate distributions and the details of comparisons between the theory and the observations summarized in table 1. Thin lines indicate the boundaries of the particular cells. Heavy lines indicate the grouping adopted in the application of the χ^2 test. Observed frequencies are written in the upper left corner of particular cells. The two other figures, each with one decimal digit, are the expected frequency (on the left) and the contribution to χ^2 of one particular cell (if the expected frequency for that cell is 3 or more) or for a group of several adjoining cells. If the expected frequencies of several cells are found to be less than 3, then they are grouped and the expected frequency is given for the entire group of cells only.

TABLE 2

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 1 vs. Cause 2, Group 1 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



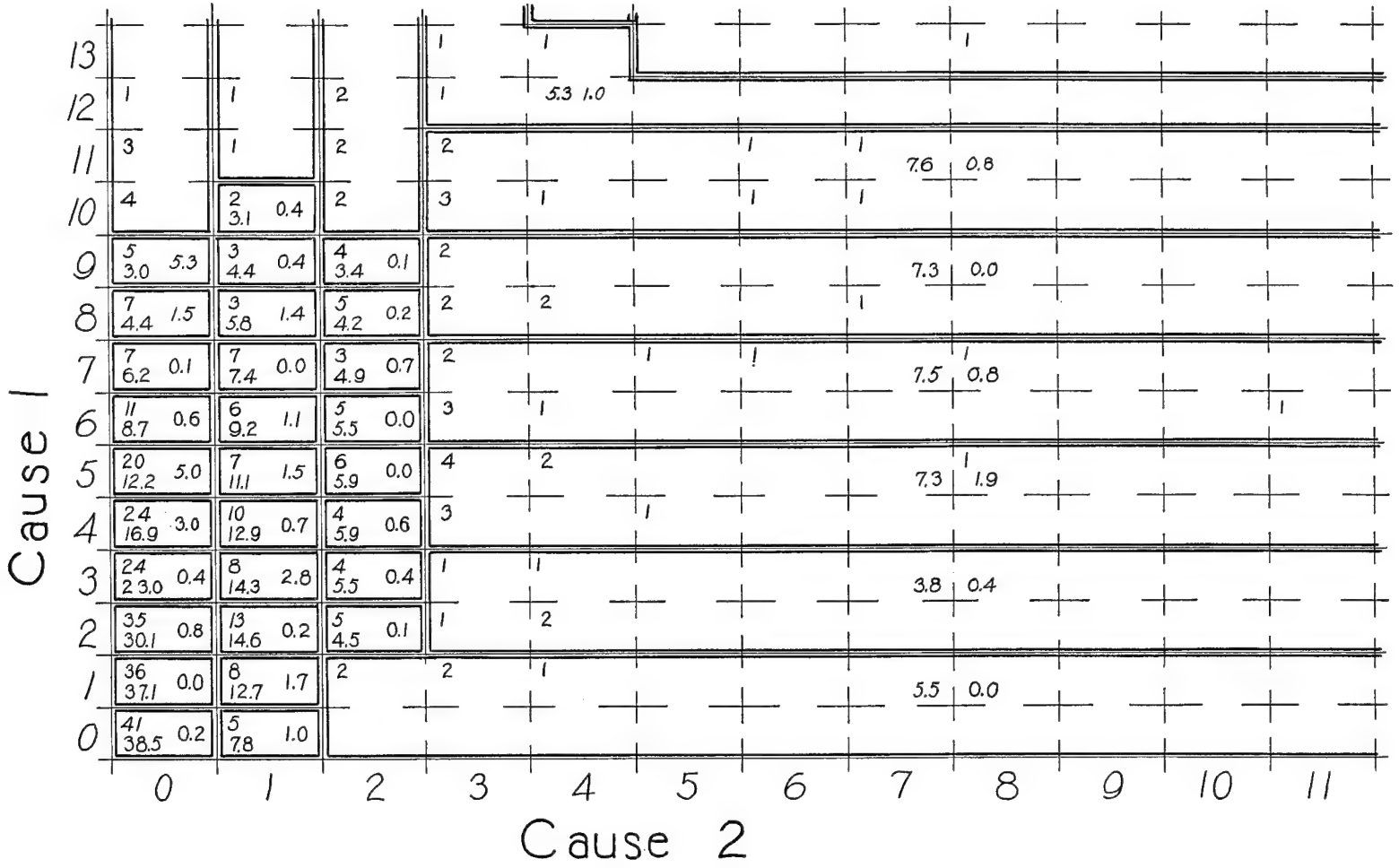
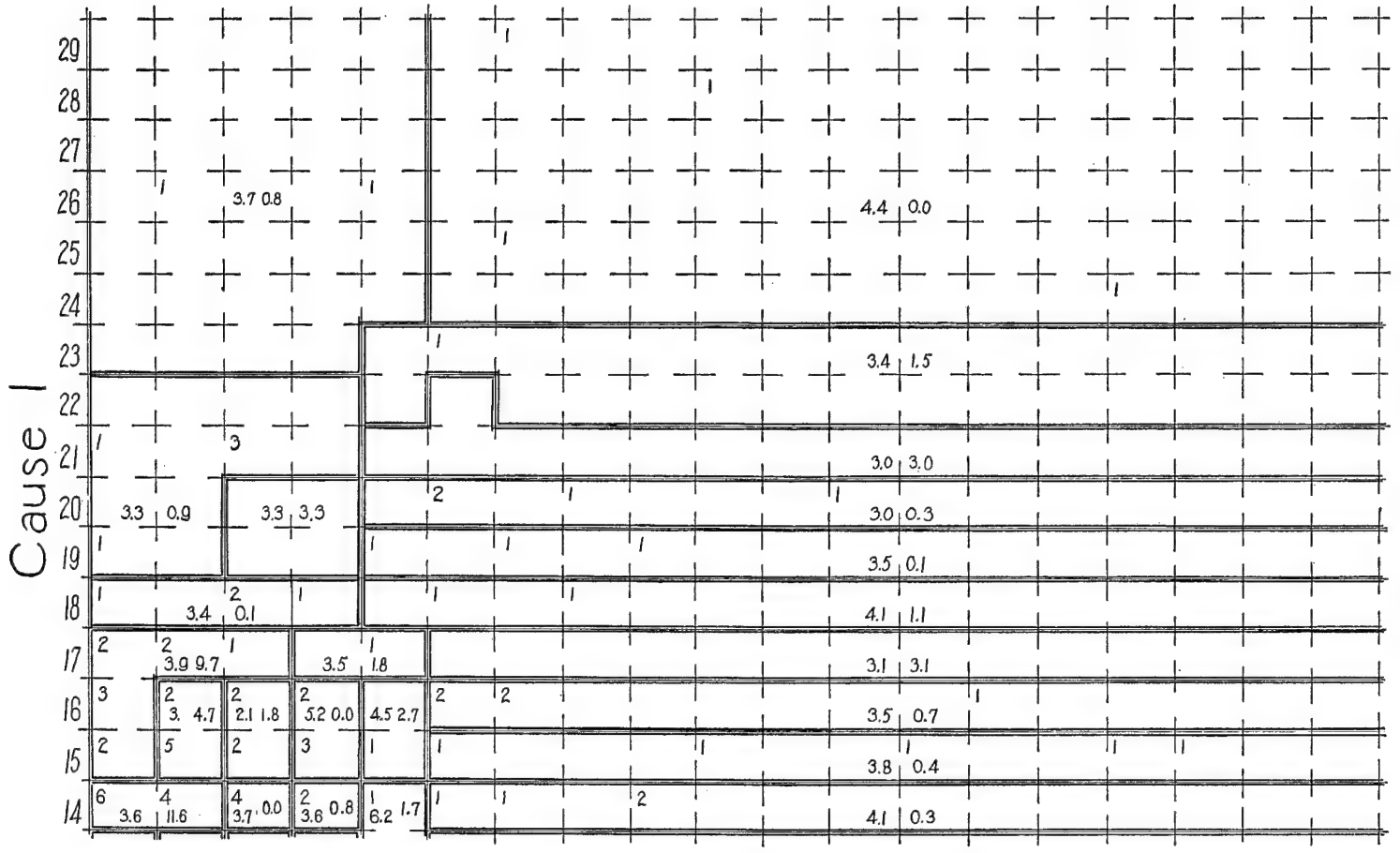

Cause 1
Cause 2
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
5.3 1.0
7.6 0.8
7.3 0.0
7.5 0.8
7.3 1.9
3.8 0.4
5.5 0.0

TABLE 3

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

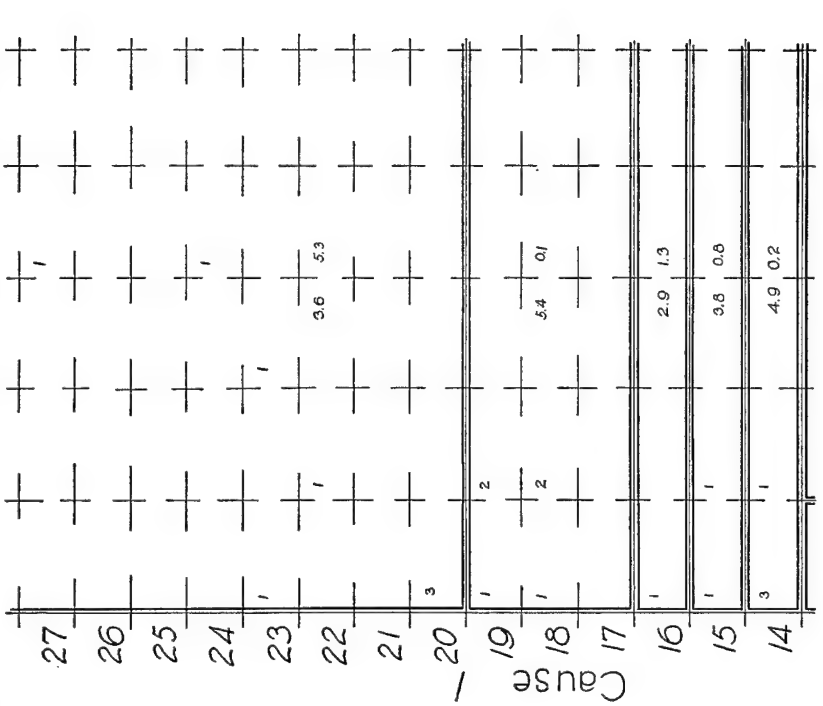
Cause 1 vs. Cause 2, Group 2 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



Cause 1
Cause 2

TABLE 4

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES
Cause 1 vs. Cause 3, Group 1 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



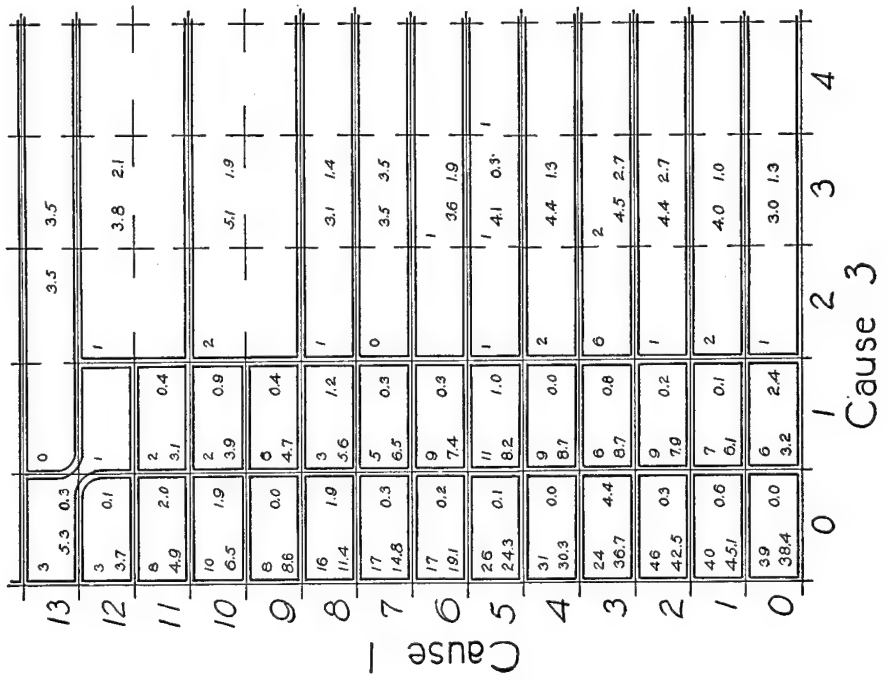

Cause 1
Cause 3
0
1
2
3
4
5
6
7
8
9
10
11
12
13

TABLE 5

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 1 vs. Cause 3, Group 2 (Div. Ind. Hyg., U.S. Pub. Health Serv.)

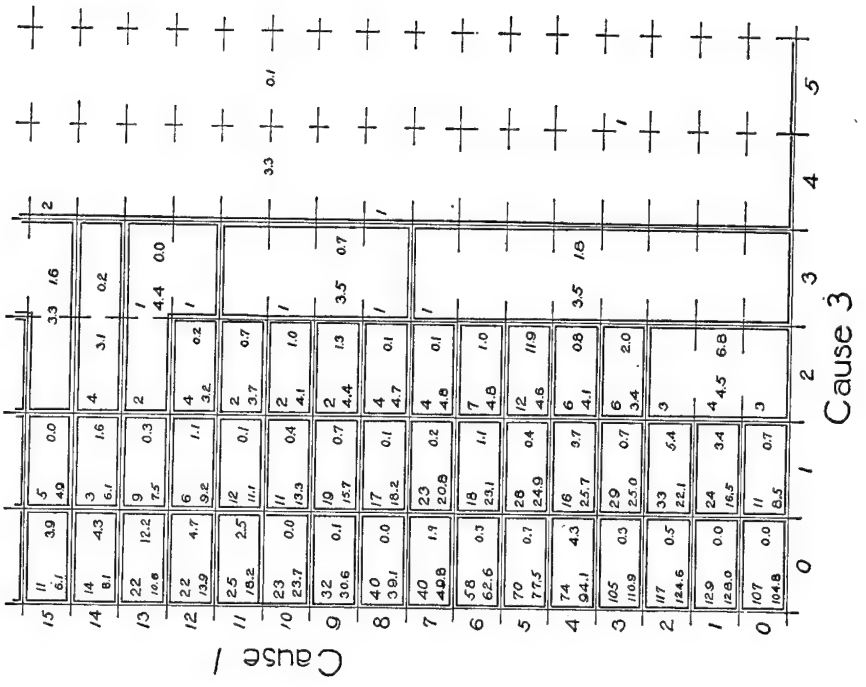

Cause 3
Cause 1

TABLE 6

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 2 vs. Cause 3, Group 1 (Div. Ind. Hyg., U.S. Pub. Health Serv.)

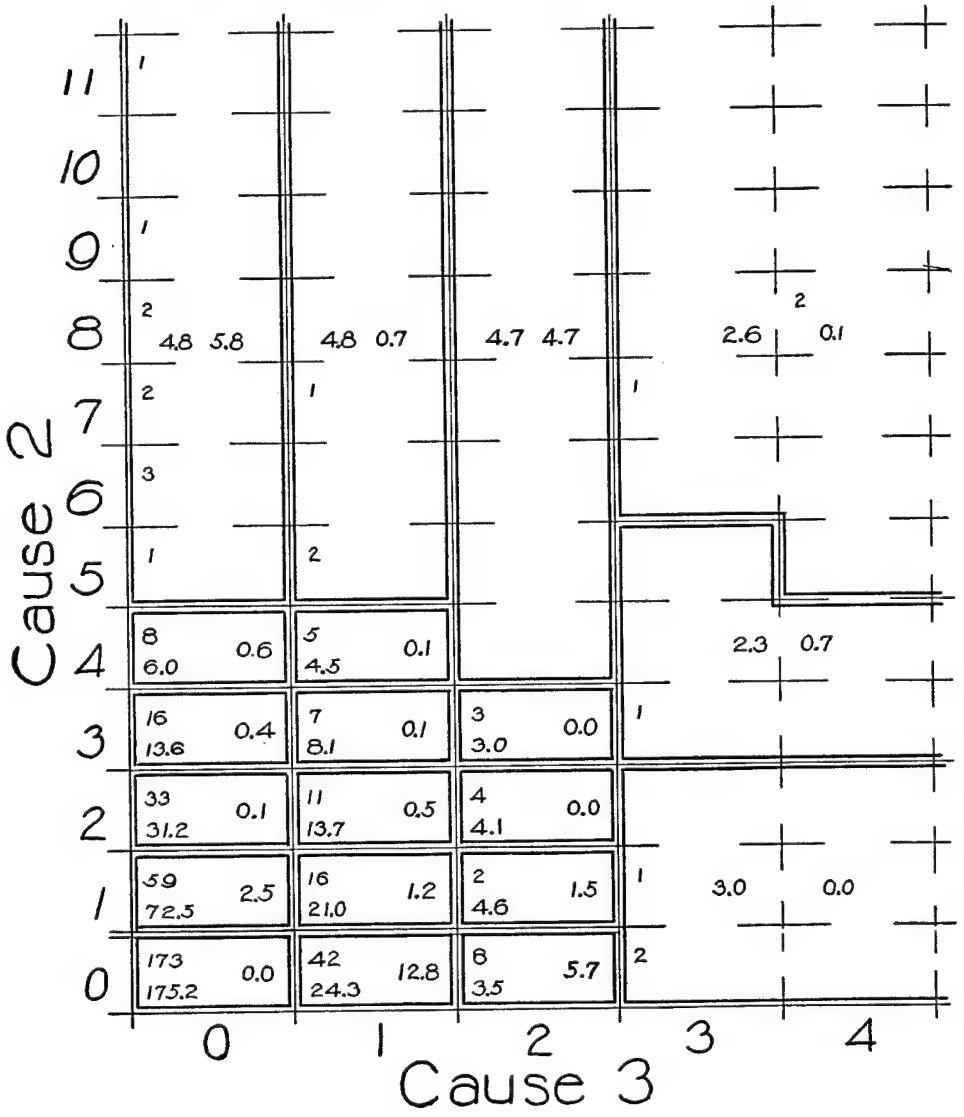
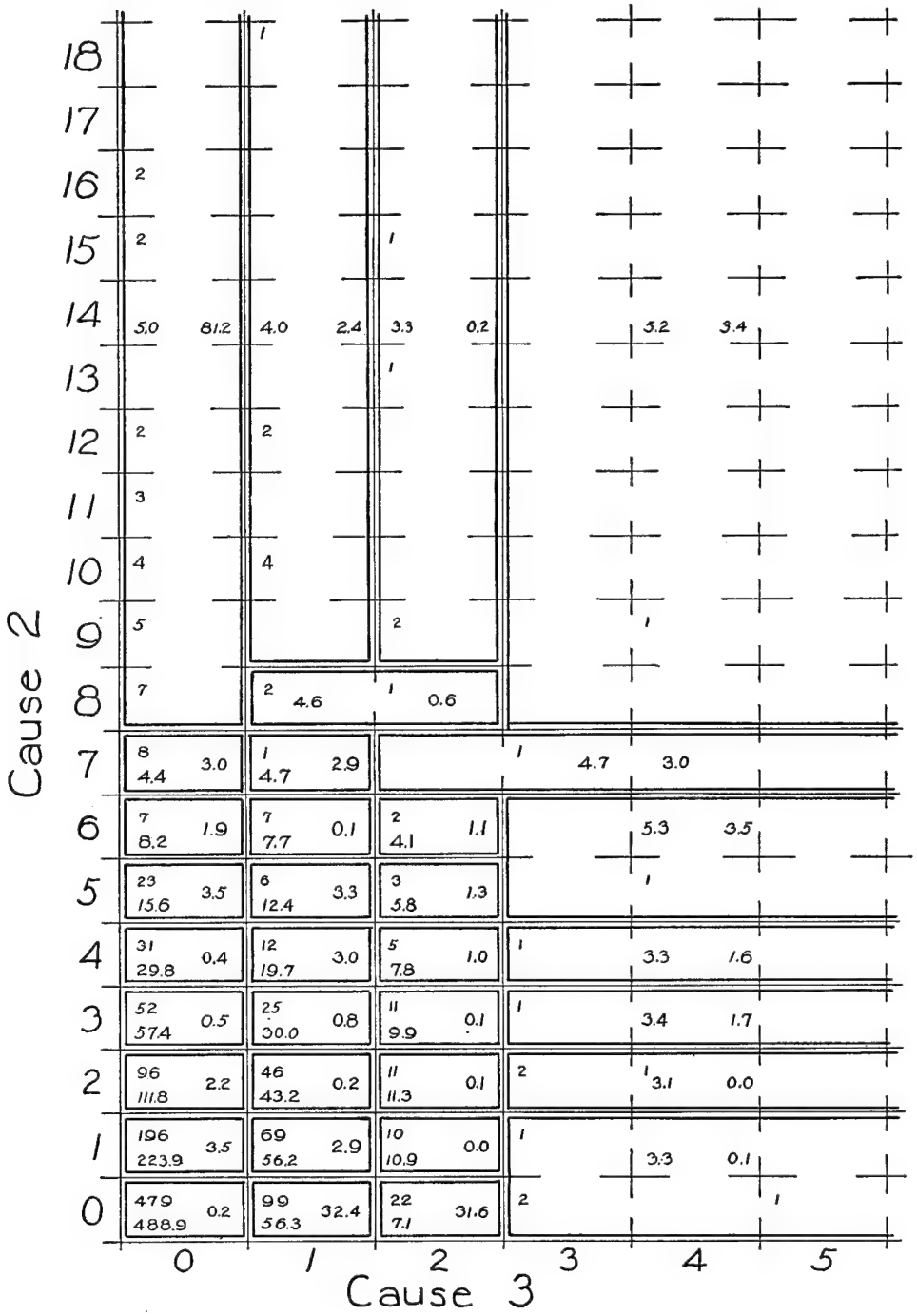


TABLE 7

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 2 vs. Cause 3, Group 2 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



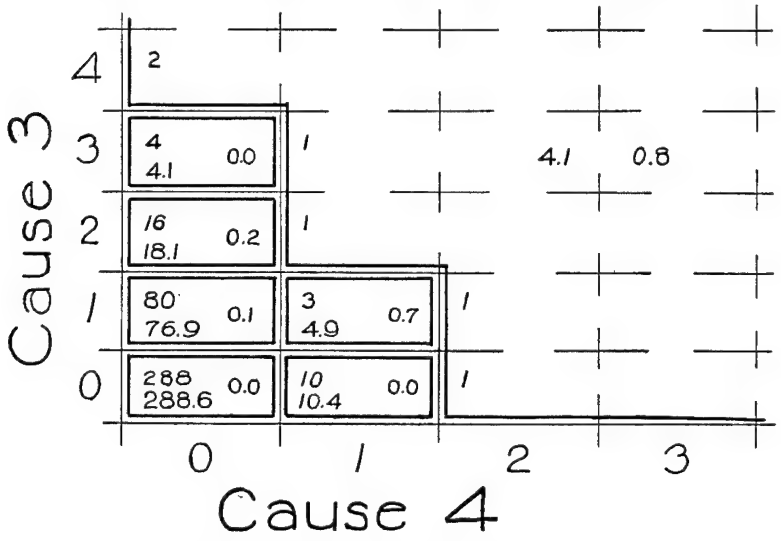
It will be seen that in three out of the ten cases studied the fit provided by the bivariate negative binomial is excellent. In two additional cases, the fit is not very good but still passable. In the remaining five cases the fit is poor.

The data summarized in table 1 refer to three groups of workers and the three samples contain 166, 407, and 1272 individuals, respectively. Cases of good and of bad fit are unevenly distributed and, in fact, in all cases relating to the largest number the fit is bad. This suggests that, probably, the true distribution of numbers

TABLE 8

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 3 vs. Cause 4, Group 1 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



of accidents does not coincide with the negative binomial in any of the cases studied. However, the divergence between the actual distribution and the negative binomial must be only slight and to detect it one needs a substantial number of observations.

Furthermore, a closer examination of tables where the fit is poor suggests that this may be due to the coexistence of two distinctly different subgroups of individuals, one large and one relatively small, with two different machineries behind the distribution of accidents. Owing to the difference in weights, the bivariate negative binomial approximates the actual distribution in the larger subgroup. However, the presence of the divergent smaller subgroup spoils the fit.

This conclusion is suggested by all the tables but the suggestion is particularly strong in the short table 9. It will be seen that the greatest contributions to the χ^2, namely 13.9 and 8.0, come from the two cells ($3 \leqq X$, $Y = 0$) and ($3 \leqq X$, $Y = 1$), with the total expected number of individuals 6.9 as against the observed 19. However, if these two cells are combined with the two corresponding cells in the same rows, the contributions of the combined cells to the χ^2 become 1.2 and 0.1 respectively and the total χ^2 sinks to a value just exceeding the 5 per cent point. Noticing that the grouping performed concerns the total of 46 individuals as against the sample of 1272, one is led to believe that, as far as the bulk of this sample is concerned, the fundamental hypothesis is not seriously wrong and that the disagree-

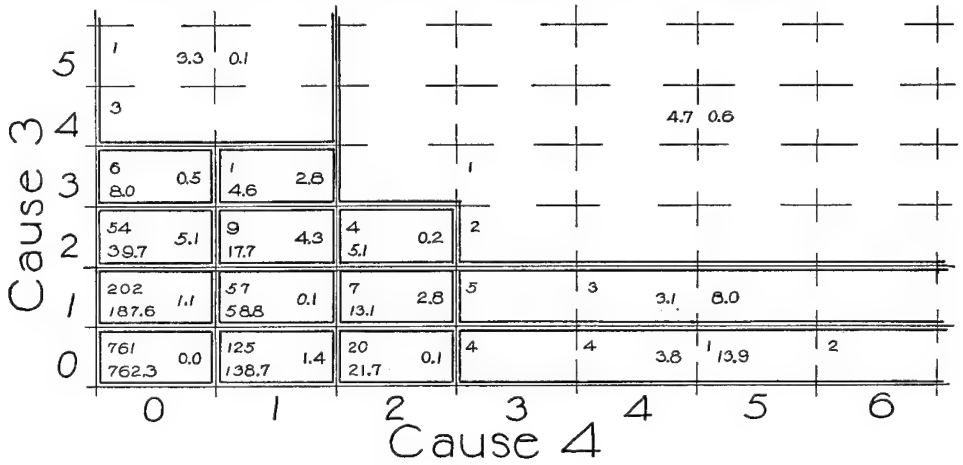
ment noted is due to a relatively small admixture of individuals with an accident proneness machinery different from that in the main body of data.

The general tentative conclusion is that cases do exist in which (*a*) the fundamental hypothesis applies approximately to accidents of two different types incurred during the same period of observation and (*b*) to the same kind of accidents incurred in two successive periods of observation. In these circumstances it is plausible that the fundamental hypothesis may be satisfied by two kinds of accidents incurred during two different periods of observation.

TABLE 9

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF INCAPACITIES

Cause 3 vs. Cause 4, Group 2 (Div. Ind. Hyg., U.S. Pub. Health Serv.)



Keeping in mind that the subject of the present paper is the possibility of using accidents of one kind to predict the number of accidents of another kind, it was thought useful to reproduce the regressions of the number of accidents of one kind on the actual number of accidents of another kind. These regressions are given in figures 1 to 5. In each the straight lines correspond to the linear equation (37) of regression based on the fundamental hypothesis.

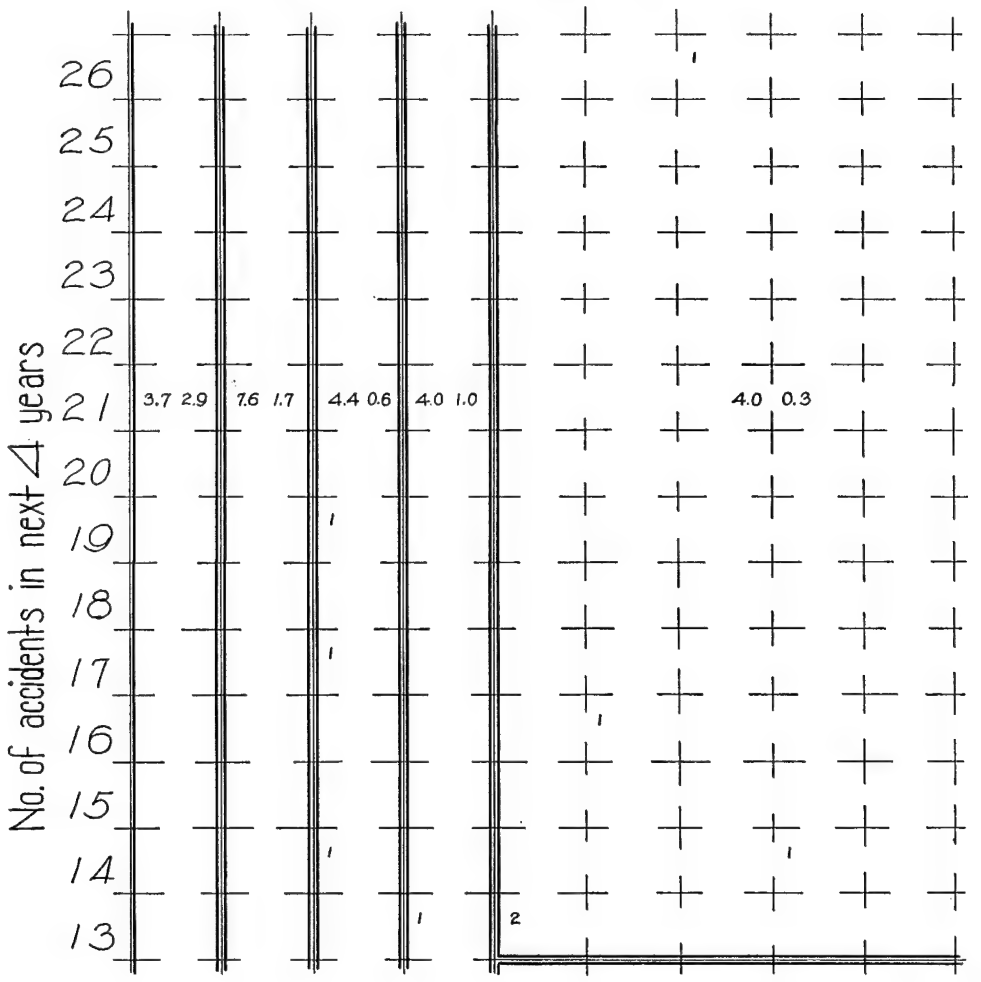
When inspecting these figures one should bear in mind that regression points corresponding to large values of the independent variable depend upon very moderate numbers of observations. Furthermore, as we have seen, the conditional variance of one variable, say Y, given a fixed value of the other, say X, increases with an increase in the value of X.

It will be seen that in many cases the fit is excellent. This is particularly true for regressions of the numbers of the less frequent accidents on those of the more frequent ones. Furthermore, the observed regression points are generally closer to the theoretical line for small values of the independent variable than for larger ones. This circumstance is important because if and when the selection of personnel is made on the ground of the number of accidents, one would naturally select those individuals who in the past had few accidents. The graphs of the regressions suggest that the results of this kind of selection will be in a reasonable agreement with predictions based on the fundamental hypothesis.

TABLE 10

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF ACCIDENTS

First Year vs. Last Four Years (Farmer and Chambers)



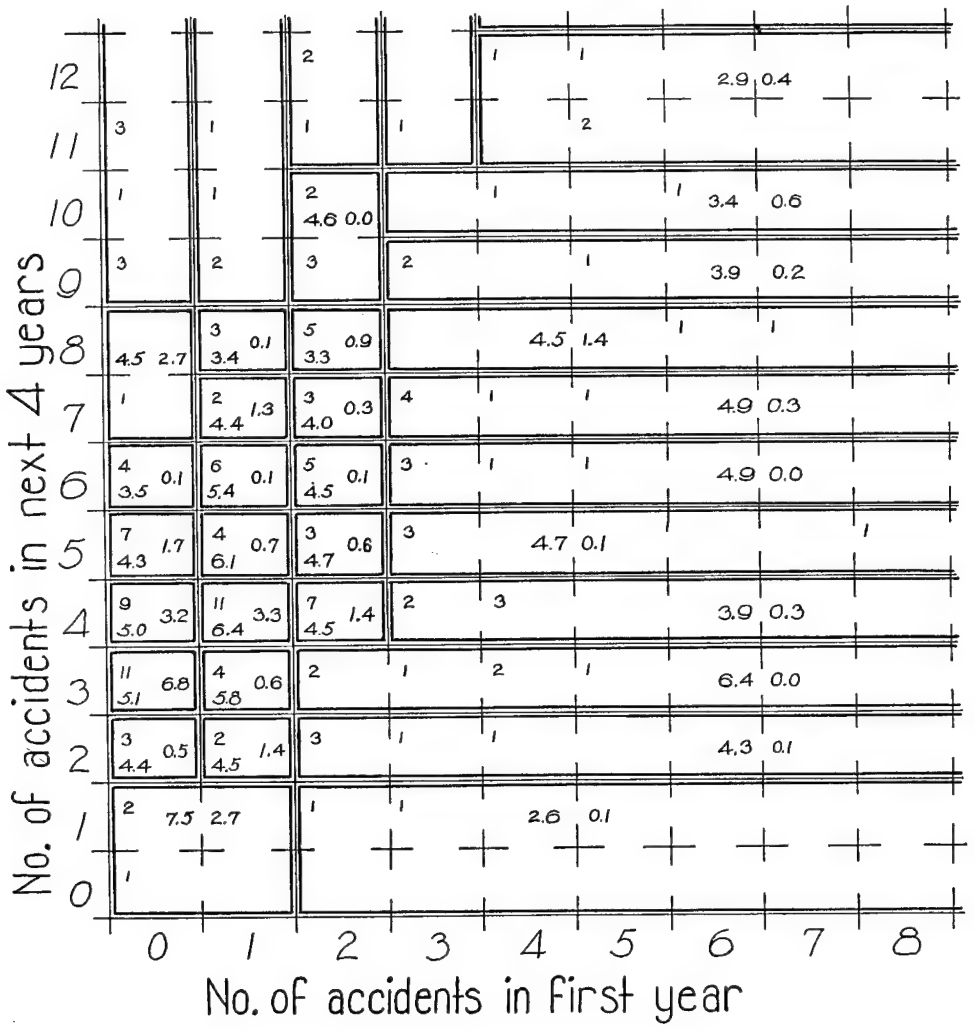
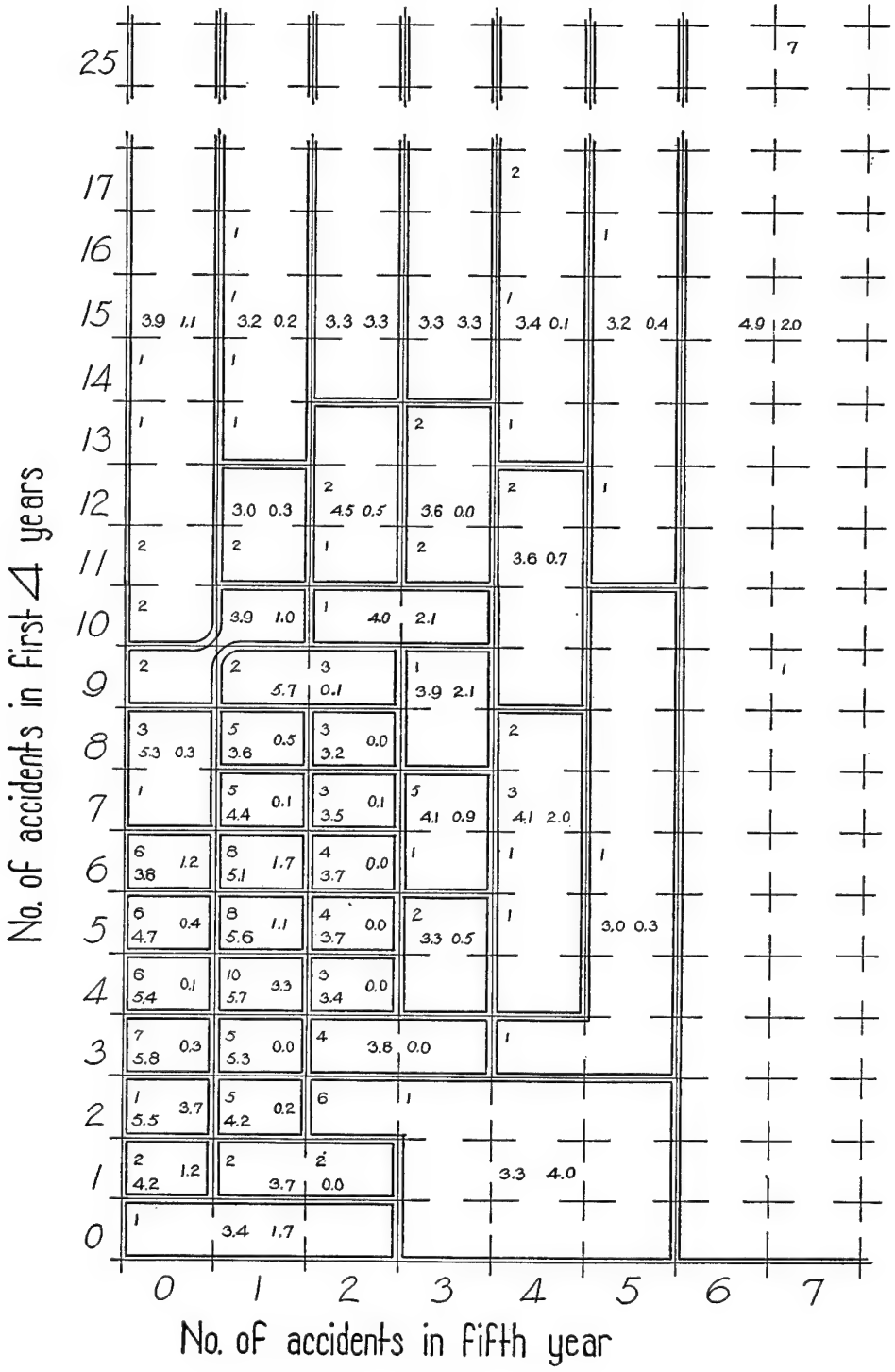

No. of accidents in next 4 years
No. of accidents in first year
0
1
2
3
4
5
6
7
8
9
10
11
12
4.5 2.7
7.5 2.7
2.9 0.4
3.4 0.6
3.9 0.2
4.5 1.4
4.9 0.3
4.9 0.0
4.7 0.1
3.9 0.3
6.4 0.0
4.3 0.1
2.6 0.1

TABLE 11

COMPARISON OF OBSERVED AND THEORETICAL DISTRIBUTIONS OF ACCIDENTS

Fifth year vs. First Four Years (Farmer and Chambers)



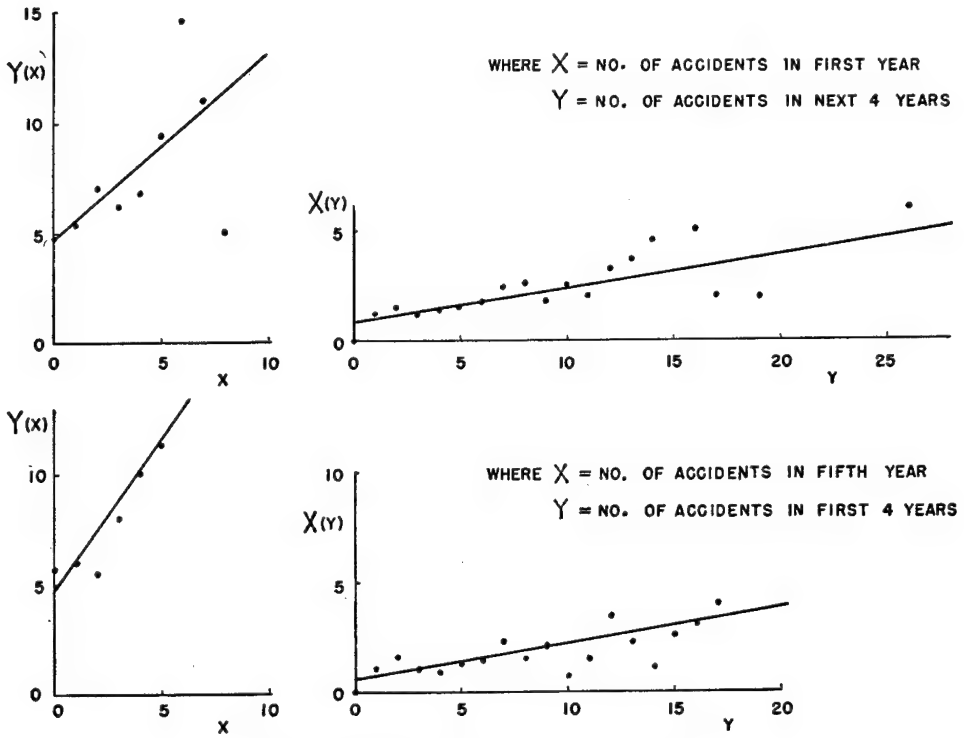


Fig. 1. Regression of X on Y and of Y on X.

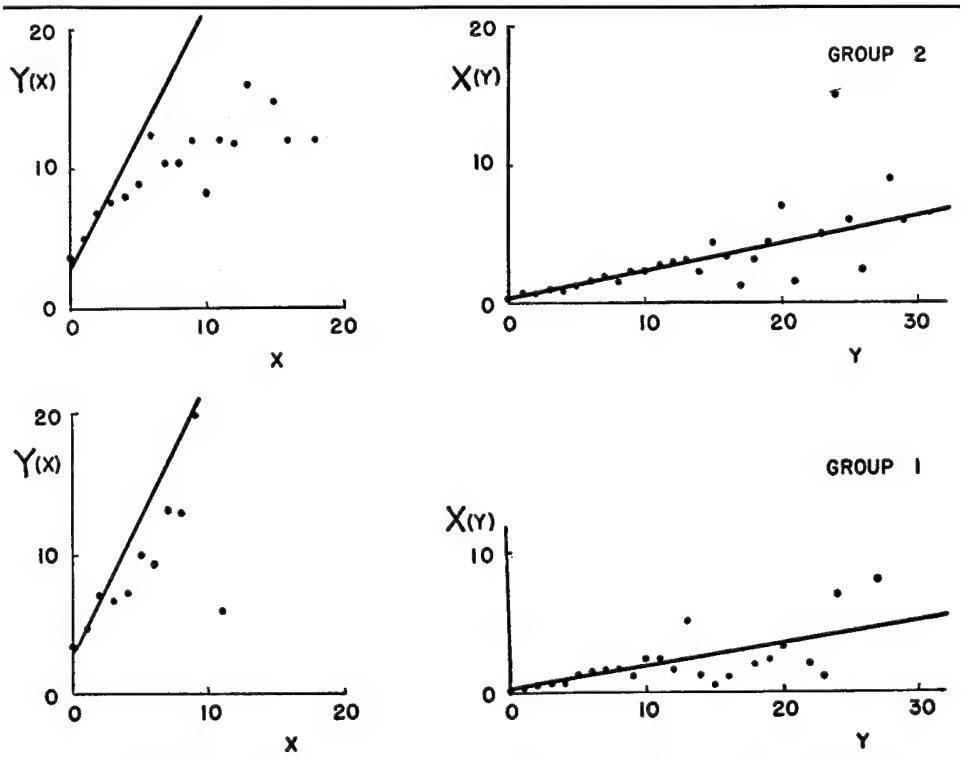


Fig. 2. Regression of X on Y and of Y on X. Where X = number of cases of digestive disease, and Y = number of cases of respiratory disease.

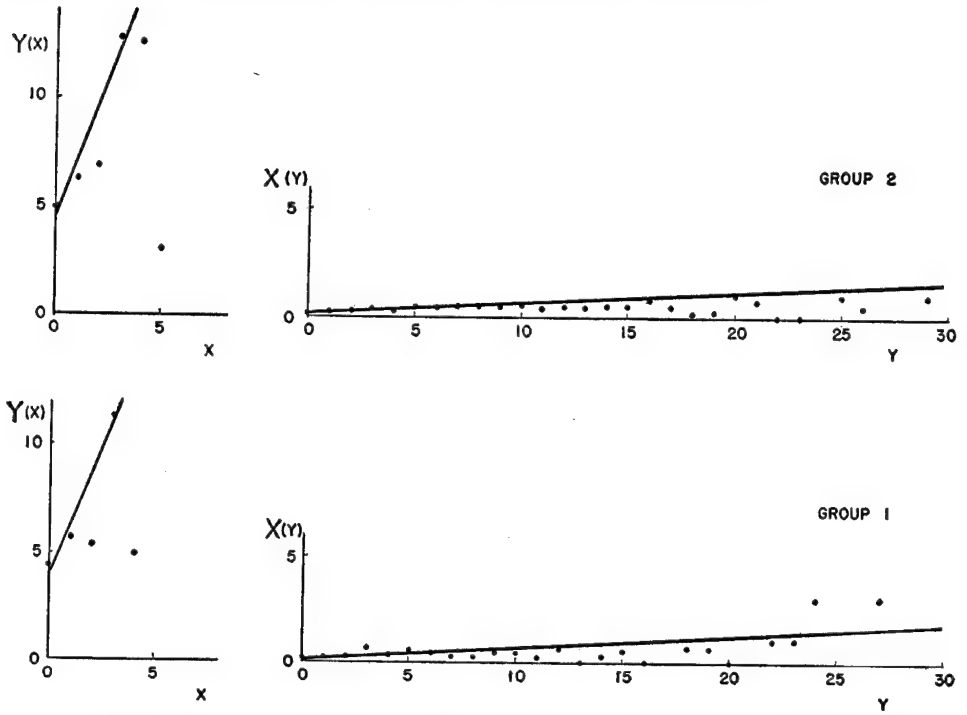


Fig. 3. Regression of X on Y and of Y on X. Where X = number of cases of nonindustrial injury, and Y = number of cases of respiratory disease.

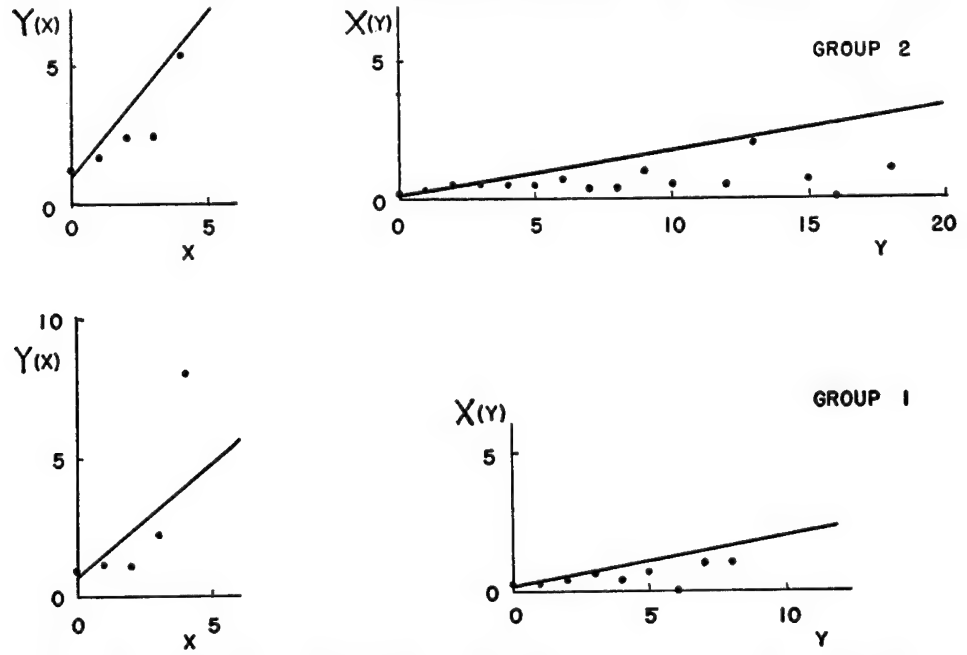


Fig. 4. Regression of X on Y and of Y on X. Where X = number of cases of nonindustrial accident, and Y = number of cases of digestive disease.

5. Measures of success in selection of personnel. In this section we study the following question. Suppose that the fundamental hypothesis applies to certain types of light and of severe accidents. Suppose further that the number Y of light accidents incurred in the past is adopted as a criterion for selecting personnel in order to diminish the number X of severe accidents to be incurred in the future. Specifically, we shall assume that the individuals selected for the particular hazardous employment will be all those for whom the number of light accidents $Y < k$

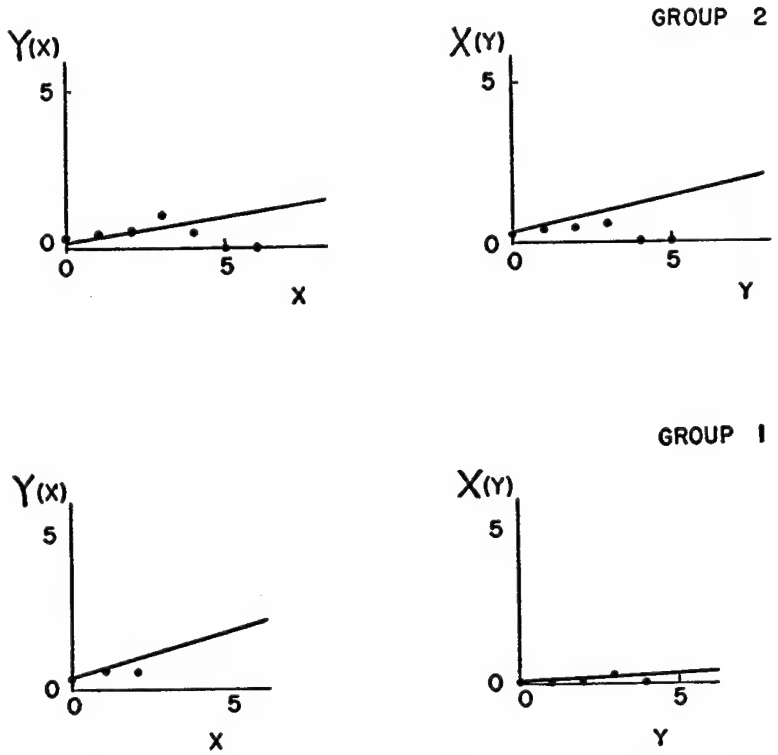


Fig. 5. Regression of X on Y and of Y on X. Where X = number of cases of industrial injury, and Y = number of cases of nonindustrial injury.

and a certain proportion Q of those for whom $Y = k$, where k and Q are so adjusted that the total number of individuals selected for employment represent a predetermined proportion P of available candidates.

In these circumstances, the interesting question is: what is the probability that in the following period of observation an individual selected for employment will have no severe accidents at all? This probability, say

$$P\{X = 0 \mid P\}\,, \tag{58}$$

compared with the probability $P\{X = 0\}$ in the unselected population, appears to be a suitable measure of the success of the selection against severe accidents.

In order to obtain $P\{X = 0\}$ we use the probability generating function (39) of X and Y and substitute in it $u = 0$ and $v = 1$. The result is

$$P\{X = 0\} = \left(\frac{\beta}{\beta + 1}\right)^{\alpha}. \tag{59}$$

This, then, is the probability of no severe accidents during the forthcoming period of observation for the nonselected population.

In order to compute (58), we first determine k and Q to satisfy the conditions imposed. The probability generating function of Y is obtained from (39) by substituting $u = 1$. Expanding the result in powers of v we get

$$p_Y(m) = \left(\frac{\beta}{\beta + A}\right)^{\alpha} \frac{\Gamma(\alpha + m)}{m!\,\Gamma(\alpha)} \left(\frac{A}{\beta + A}\right)^{m}. \tag{60}$$

The number k is determined by the condition

$$\sum_{m=0}^{k-1} p_Y(m) \leqq P < \sum_{m=0}^{k} p_Y(m). \tag{61}$$

Then

$$Q = P - \sum_{m=0}^{k-1} p_Y(m). \tag{62}$$

Once k and Q are found, then (58) is computed by a simple application of the formula of Bayes with the use of (40).

$$\begin{aligned} P\{X=0|P\} &= \frac{1}{P}\left[P\{(X = 0)\,(Y < k)\} + Q\,\frac{P\{(X = 0)\,(Y = k)\}}{P\{Y = k\}}\right] \\ &= \frac{1}{P}\left[\left(\frac{\beta}{\beta+A+1}\right)^{\alpha} \sum_{m=0}^{k-1} \frac{\Gamma(\alpha+m)}{m!\,\Gamma(\alpha)} \left(\frac{A}{\beta+A+1}\right)^{m} + Q\left(\frac{\beta+A}{\beta+A+1}\right)^{\alpha+k}\right]. \end{aligned} \tag{63}$$

Suppose that for a given population of candidates for employment and for a given pair of kinds of accidents the values of α, β and A have been determined. Suppose further that the proportion P of candidates to be selected for employment is also determined. In order to estimate the prospective success of selection of candidates we first compute the standard of comparison (59) and then determine k and Q to satisfy (61) and (62). Then these values are substituted into (63).

Naturally, the effect of selection of candidates depends on all four parameters involved, on α and β characterizing the distribution of Λ in the population of candidates for employment, on the number A and on the proportion P of those to be selected. In the unselected population the expectation of Λ and its variance are

$$E(\Lambda) = \frac{\alpha}{\beta}, \qquad \sigma_{\Lambda}^2 = \frac{\alpha}{\beta^2} = \frac{E(\Lambda)}{\beta}. \tag{64}$$

If the variance σ_Λ^2 is very small—and this will happen when β is a larger number—then even a very sharp selection will give practically no result. In the cases considered in table 1 the values of β are moderate and, therefore, the prospects for selection are promising. Turning to the other factors involved, it must be obvious that the smaller P is the sharper must be the selection and, therefore, the greater its effect. Finally, the effect of selection depends considerably on the value of A, which is the ratio of the average frequencies of light and of severe accidents,

$$A = \frac{E(Y)}{E(X)}, \tag{65}$$

in the unselected population. Because of this interpretation the quotient A may be called the modulus of the relative frequency of light accidents.

TABLE 12

CORRESPONDING VALUES OF k AND Q FOR A SET OF INCREASING VALUES OF THE MODULUS OF RELATIVE FREQUENCY A

A	$\alpha = 3, \beta = 2$				$\alpha = 3, \beta = 1$			
	$P = .125$		$P = .250$		$P = .125$		$P = .250$	
	k	Q	k	Q	k	Q	k	Q
1	0	.125	0	.250	1	0.0	1	.125
2	1	0.0	1	.125	2	.0139	3	.0401
3	1	.061	2	.0708	3	.0215	5	.0065
4	2	.0139	3	.0401	4	.0261	6	.047
5	2	.0517	4	.0203	5	.0292	8	.025
10	5	.0292	8	.0252	11	.0174	17	.0038
15	8	.0220	12	.0282	17	.013	25	.014
20	11	.0172	17	.0038	24	.0004	34	.0054

The actual numbers characterizing the possible effect of selection are of practical importance. With this in mind table 12 and figures 6 and 7 were constructed. They illustrate two hypothetical situations. In one of them the values of $\alpha = 3$ and $\beta = 2$ approximately coincide with those corresponding to the experience of the London bus drivers (see table 1). In the other case, $\alpha = 3$ and $\beta = 1$, so that both the expectation of Λ and its variance are increased. The figures are intended to illustrate the effect of selection corresponding to two different levels of sharpness of selection. In one case we assume $P = .125$ and in the other $P = .250$. The value of the modulus A varies from $A = 1$ to $A = 20$. For a succession of increasing values of A, table 12 gives the corresponding values of k and Q with which the proportion of selected candidates will be equal to P. Figures 6 and 7 give the corresponding values of $P\{X = 0|P\}$. The horizontal dashed line indicates the standard of comparison $P\{X = 0\}$. It is seen that in both cases, when A is small, the effect of selection is already noticeable. When A is substantial, say $A \geqq 5$, then the probability of avoiding severe accidents is considerably increased by selection. The practical con-

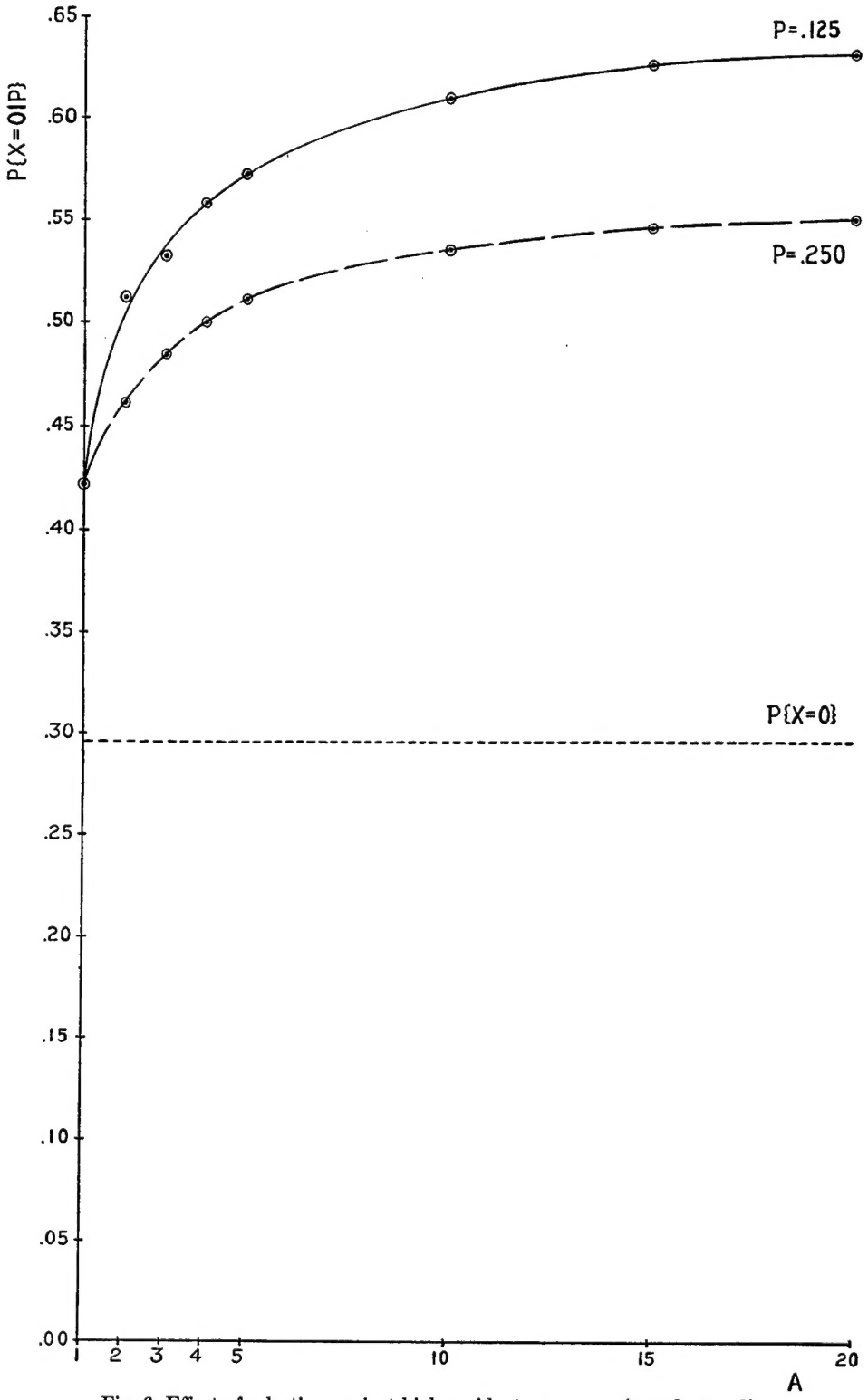


Fig. 6. Effect of selection against high accident proneness ($\alpha = 3$, $\beta = 2$).

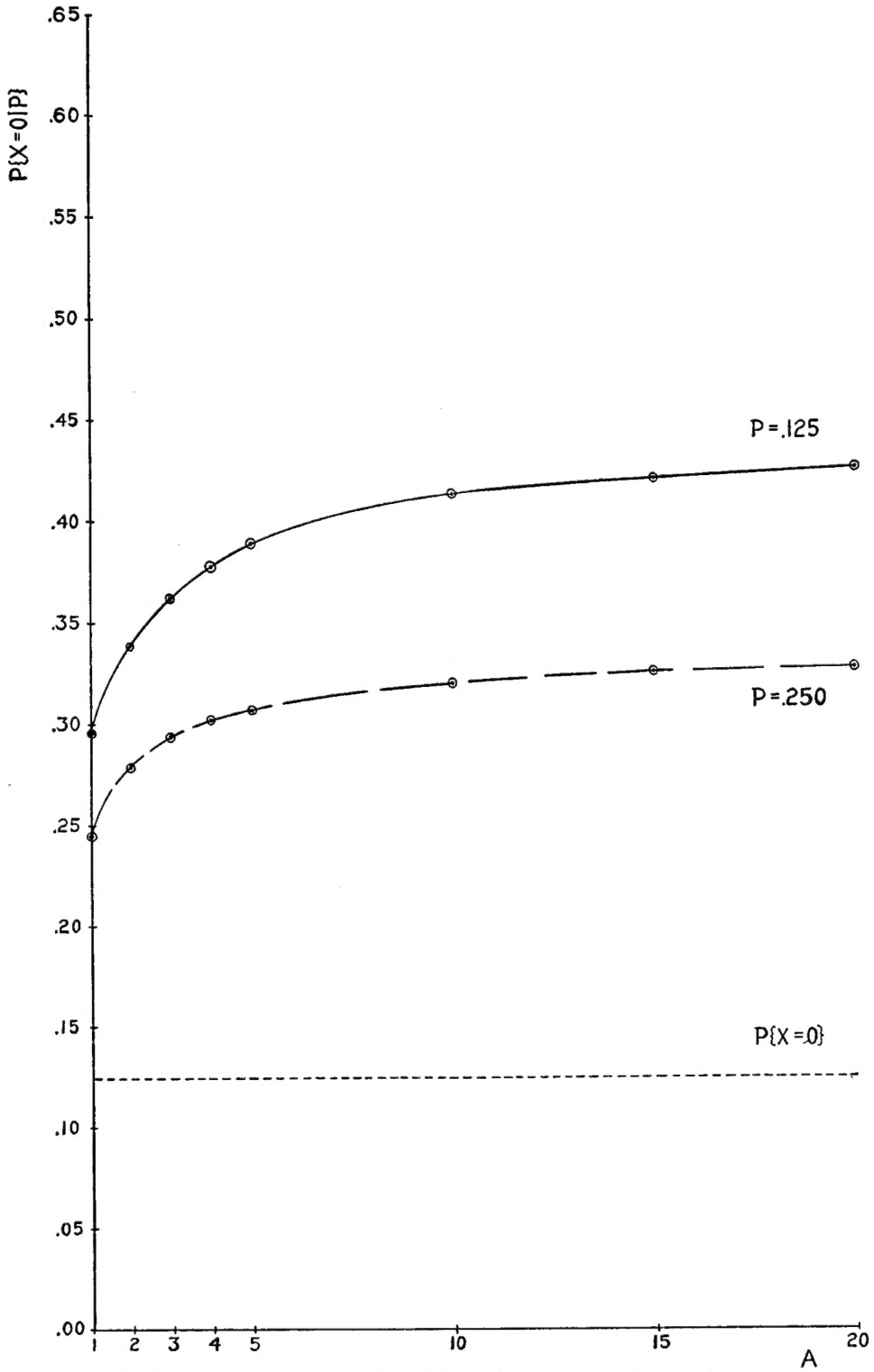


Fig. 7. Effect of selection against high accident proneness ($\alpha = 3$, $\beta = 1$).

clusion suggested by this result is that, in order that the selection of personnel on the basis of light accidents incurred in the past be successful, it is desirable that the average number of light accidents during the period of observation be large. This may be achieved either by taking a long period of observation (which may be impracticable) or by using some artifice to increase the exposure to light accidents during a relatively short period of observation.

6. Joint distribution of the number of light accidents and of the number of survived severe accidents. As mentioned in the introduction, even if the fundamental hypothesis assumed in this paper is strictly satisfied with regard to a category of light accidents and a category of severe accidents, if these latter accidents are really severe, then their number incurred during a fixed period of time will not follow the negative binomial distribution. The reason is that from time to time a severe accident, occurring at the early part of the period of observation, will prove fatal to the individual concerned. As a result, there will be no exposure of this individual to possible further severe accidents during the same period of observation. Thus, if and when statistics relating to light and to severe accidents sustained by the same individual become available, then in order to be able to verify the fundamental hypothesis and to estimate the constants involved, a new type of distribution will be necessary. This must take into account the fact that each severe accident may lead to invalidism or to death for the individual concerned. The purpose of this section is to consider this distribution. Our basic assumption, supplementing the fundamental hypothesis, will be that each individual involved in a severe accident has the same probability θ of surviving the accident and continuing the employment with all its hazards. The alternative to such survival will be either death or retirement from the particular employment. However, this distinction may be ignored and we shall speak of two possibilities only: survival (in good health) or death (the latter meaning either actual death or retirement).

In connection with the change in the problem, we shall need new notation. The letter Y will be used, as formerly, to denote the number of light accidents incurred by an individual during a period of observation. On the other hand, the letter X will be used to denote the number of severe accidents *that this individual will survive*, incurred by the individual during the same or a different period of observation. Thus, if an individual incurs three severe accidents and dies at the third, then for this individual $X = 2$. In order to distinguish between deaths and survivals we shall need a third random variable Z. This variable will be defined to be equal to zero if the particular individual survives all the period of observation, and unity if the individual does not.

The statistics of light and severe accidents may be divided into two categories. First we postulate the availability of the numbers of light and of severe accidents for those individuals who survived the entire period of observation of severe accidents. The figures obtainable from these statistics will be the empirical counterpart of the theoretical probabilities $P\{(X = k)\ (Y = m) \mid Z = 0\}$. The second part of the statistics contemplated would refer to individuals who died as a result of a severe accident during the period of observation. The figures obtainable from such statistics would correspond to probabilities $P\{(X = k)\ (Y = m) \mid Z = 1\}$. The formulae for the probability generating functions for these relative probabilities arise as limiting

forms of generating functions deduced under the general hypotheses considered in Part II of this paper and are as follows:

$$G_{X,\,Y|Z=0}(u,\,v) = \left(\frac{\beta + 1 - \theta}{\beta + 1 - \theta u + A(1 - v)}\right)^{\alpha}, \tag{66}$$

$$G_{X,\,Y|Z=1}(u,\,v) = \frac{(\beta+1-\theta)^{\alpha}}{(\beta+1-\theta)^{\alpha}-\beta^{\alpha}} \cdot \frac{1-\theta}{1-\theta u}\left[\frac{\beta^{\alpha}}{[\beta+A(1-v)]^{\alpha}} - \frac{\beta^{\alpha}}{[\beta+1-\theta u+A(1-v)]^{\alpha}}\right]. \tag{67}$$

It is seen that for individuals who survive the period of observation of severe accidents the joint distribution of the number Y of light accidents and of the number X of survived severe accidents is again a bivariate negative binomial. On the other hand, for individuals who die as a result of a severe accident, the joint distribution of X and Y is more complicated, with probability generating function given by formula (67).

If and when the data on light and severe accidents are available, formulae (66) and (67) could be used to test the validity of the fundamental hypothesis assumed in the present paper.

REFERENCES

[1] M. Greenwood and G. U. Yule, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attack of disease or of repeated accidents," *J. Roy. Stat. Soc.*, Vol. 83 (1920), pp. 255–279.

[2] E. M. Newbold, *A contribution to the study of the human factor in the causation of accidents.* Industrial Health Research Board, Report No. 34. London, H. M. Stationery Office, 1926.

[3] E. Farmer and E. G. Chambers, *A study of accident proneness among motor drivers.* Industrial Health Research Board, Report No. 84. London, H. M. Stationery Office, 1939.

[4] Ove Lundberg, *On Random Processes and Their Application to Sickness and Accident Statistics.* Uppsala, Almquist and Wiksells, 1940. 172 pp.

[5] G. Pólya, "Sur quelques points de la théorie des probabilités," *Ann. de l'Institut Henri Poincaré*, Vol. 1 (1930), pp. 117–161.

[6] William Feller, "On the theory of stochastic processes, with particular reference to applications," *Proceedings, Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, 1949, pp. 403–432.

[7] J. Neyman, *First Course in Probability and Statistics.* New York, Holt, 1950.

[8] J. Neyman, "Contribution to the theory of the χ^2 test," *Proceedings, Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1949, pp. 239–275.

[9] F. Y. Edgeworth, "On the probable errors of frequency constants," *Journ. Roy. Stat. Soc.*, Vol. 71 (1908), pp. 662–678 (Appendix).

[10] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc.*, Ser. A, Vol. 222 (1922), pp. 309–368.

[11] Evelyn Fix and Jerzy Neyman, "A simple stochastic model of recovery, relapse, death and loss of patients," *Human Biology*, Vol. 23 (1951), pp. 205–241.